Anomaly Detection and Accuracy Measurement for Categorical Data

Kameron Grubaugh, Zachary Zimmerman, Nicholas McAfee, Emily McGowan, and Paul Evangelista

Department of Systems Engineering
United States Military Academy
West Point, NY 10996, USA

Corresponding author's email: Grubaugh.Kameron@aol.com

Author Note: The authors participated in a year-long senior research project in the Department of Systems Engineering at the United States Military Academy under the advisement of Paul Evangelista. Paul Evangelista is a colonel in the U.S. Army and currently serves as an Academy Professor.

Abstract: The Department of Defense (DoD) recently initiated an effort to compile all inter-service maintenance data for equipment and infrastructure, requiring the consolidation of maintenance records from over 40 different data sources. This research evaluates and improves the accuracy of this maintenance data warehouse by means of value modeling and statistical methods for anomaly detection. The first step in this work included the categorization of error-identifying metadata, which was then consolidated into a weighted scoring model. The most novel aspect of the work involved error identification processes using conditional probability combinations and likelihood measures. This analysis showed promising results, successfully identifying numerous invalid maintenance description labels through the use of conditional probability tests. This process has potential to both reduce the amount of manual labor necessary to clean the DoD maintenance data records and provide better fidelity on DoD maintenance activities.

Keywords: Corrosion, Anomaly Detection, Data Analysis, Value Modeling, Hash Data Structures, Categorical Data

1. Introduction

Corrosion of equipment and weapons systems accounts for significant cost within the Department of Defense (DoD) maintenance budget and results in unavailability of mission-essential resources. To combat this problem, Logistics Management Institute (LMI) compiles a consolidated data warehouse with records from various military services and primarily attempts to analyze this data using an action-object classification method to recommend action for key DoD stakeholders. However, the action-object identification process does not always yield correct associations and there are no systems in place to evaluate the performance of these predictive models across the entire database. This paper addresses common problems with metadata and large databases by implementing a scorecard model to assess accuracy and leveraging anomaly detection to identify errors.

1.1 Background

The DoD Operation and Maintenance (O&M) budget accounts for one of the most substantial and fastest growing portions of military expenditure. With regard to depot maintenance alone, the FY 2019 budget request reflects an increase of nearly $2.4 billion in program growth (Office of the Under Secretary of Defense, 2018). Falling largely under this account, corrosion continues to cost the military large allocations of its budget as well as time in downed equipment. By considering how cost structure and financial data affect the military's management of its resources, we seek to gain insight into the importance of properly reporting corrosion data as it relates to policy decisions on an organizational level. Data itself does us no good if policymakers cannot pull relevant and lucid conclusions from aggregate records. According to existing government literature and analysts' findings, the Department of Defense could benefit from an increased urgency on data organization, greater emphasis on life cycle costing, and the increased use of mathematical models in analyzing best courses of action with regard to preventative and corrective corrosion maintenance.

Ultimately, a more long-term focus on systems' cost, to include those due to corrosion, would better conserve resources than acquiring cheaper alternatives in the present. Bhaskaran, Palaniswamy, Rengaswamy, & Jayachandran (2005) consider life cycle costing the most effective of the methods currently in use in corrosion analysis, as it acts as a framework that seeks to minimize the cost by determining the annualized value of each corrosion control option. The Office of the Secretary of Defense (2014) has even published guidance requiring that the military update O&M cost estimates during multiple stages of the life cycle of major weapons systems. This builds off of DoD Instruction 5000.67, which provides guidance on the prevention and mitigation of corrosion, placing emphasis on accounting for life-cycle cost when making acquisition decisions (Office of the Under Secretary of Defense, 2010). These directives represent a strategic shift in how the military acquires systems. Under the old focus, as the Defense Science Board suggested as early as 2004, the system incentivizes program managers to minimize acquisition cost rather than life-cycle cost (Office of the Under Secretary of Defense, 2010). The Defense Acquisitions University has since increased its emphasis on life-cycle cost as a criterion for acquiring new systems; however, the collection of information regarding corrosion cost still requires improvement.

1.2 Related Work

Before developing our own analysis methods, we referenced work previously done in pattern recognition, big data analytics, data mining, and anomaly detection. This report will focus on efforts to model an accuracy measure for the client and later describe an anomaly detection method to find outliers in the data within the Maintenance Availability Data Warehouse (MADW). Ultimately, this will help find ways to improve the overall accuracy of the action-object prediction process.

Many industries and organizations apply text and data mining techniques to effectively analyze big data. Focusing mainly on techniques that are used for ‘unstructured data,’ these methods include, but are not limited to, text analytics, information extraction, text summarization, question answering, and sentiment analysis (Gandomi and Haider, 2015). Each of these have slightly different approaches but accomplish similar objectives – using big data to enhance decision making. Current realms of application for these methods include business, technology, healthcare, and tourism. Additionally, data analysis can allow businesses to give personalization to customers and discover patterns in their operations (Yung, 2015). We focus here on anomaly detection as a means most applicable to corrosion maintenance, which could affect progress in corrosion prevention. One such method to do this is the use of unsupervised anomaly detection, or the detection of records that do not adhere to expected normality. In order to define normality for the given dataset, which includes information on rotary wing aircraft for the purposes of this study, we use conditional probability tests such as the suspicious coincidence measure and Bayes Theorem (Barlow, 1989; Bayes, 1763). The suspicious coincidence method finds the ratio that satisfies $r = \frac{P(A \cap B)}{P(A)P(B)}$ where a low r value signifies that an event does not occur naturally (Das and Schneider, 2007). Similarly, Bayes Theorem will provide the probability of an event occurring given a condition, following the equation $P(A|B) = \frac{P(A \cap B)}{P(B)}$, in which a low value signifies a lower chance that the event occurs naturally. This methodology can expand to include multiple dimensions of analysis beyond the two depicted in these equations.

The main predictive work that LMI performs on this data warehouse is the categorization of maintenance actions and objects (often referred to as end items). To do this accurately, organizations currently implement a variety of pattern recognition methods. We identify the e-commerce approach as most applicable to the corrosion datasets (Shen, Ruvini, & Sarwar, 2012). In an industry that must display thousands of items in specific, user friendly and easily readable categories, categorization is incredibly important. Taking a hierarchical approach, the top layer of categorization is similar to the department you wish to shop in when visiting a retail site; for the MADW it is the end item, such as engine or rotary wing aircraft. Following this approach down to the lowest level, a single item a consumer wishes to purchase could be a member of four sub-categories to the department, much like a specific record in the corrosion data. The problem many metadata users face compounds when entries are incorrect because there are often no frustrated consumers sending complaints about the incorrectly categorized items. Instead, workers must sort through the hundreds of millions of entries and identify the errors; if the errors go uncaught, it affects the accuracy of the dataset. For example, if an analyst noticed the object *parachute* paired with the system *engine*, something is clearly wrong. To notice this in a larger scale and avoid using thousands of man-hours to find these anomalies, a detection algorithm can identify those entries that do not fit the normal or frequently used combinations. Metadata users could use this to search for only those records and save a significant quantity of labor-hours and dollars.

2. Methods

We developed two methods for this research: (i) a metric for determining the value of the MADW based on accuracy and utility in respect to the overarching goal of corrosion analysis, and (ii) an anomaly detection method to improve the error identification process in a large dataset. The two objectives, while requiring different approaches, build off of one another, as the initial work on scorecard measures helps focus the later attempts at automation of error detection methods.

2.1 System Evaluation Process

In order to evaluate aggregate accuracy of the data warehouse, we relied on a weighted value model based on categorizing data into several functional categories and assigning scores in accordance with stakeholder expectations. When compiled, the result is a scorecard value which measures the usefulness of the dataset in regard to the primary goal of the study. The metadata used for this analysis consisted of a variety of prefabricated error checks, each denoting the number of records in the warehouse that fit a particular inconsistency that LMI identified as erroneous. We then considered all of these checks and grouped them into seven categories considering their effect on database usefulness, fundamentally composed of relevance to object-action recognition, system availability, and costing. Each category represents a simple sum of all the erroneous records, and each associated variable x_i is a ratio of one minus that aggregate over the total number of checked records. The higher the value of x_i, the more accurate the records for a given field. The following equation measures the value of the variable:

$$x_i = 1 - \frac{\sum \text{erroneous records attributed to category } i}{\sum \text{total records in the dataset}} \tag{1}$$

where *i* represents one of seven possible categories: end item (1), duration severe (2), action (3), cost severe (4), duration minor (5), cost minor (6), and miscellaneous (7). Note that this does present the possibility of double counting erroneous records, if multiple fields in those records were flagged in different checks; however, based on our data exploration we do not view this as a large enough threat to diminish the utility of the model. Next, we interviewed our primary stakeholder to develop reasonable weights for each category in terms of the creation of an overall scorecard value. We then validated the categorization of fields into the categories represented by the variables. As expected, the checks relating to object accuracy were most crucial to the scorecard value of the data warehouse. The final expression for calculating a scorecard value is presented here:

$$0.5x_1+0.125x_2+0.1x_3+0.085x_4+0.075x_5+0.075x_6+0.04x_7 \tag{2}$$

2.2 Anomaly Detection Using Conditional Probability

We approached the second phase of this research with the assumption that some erroneous records would exhibit anomalous behavior, and thus anomaly detection would serve as an additional technique to improve accuracy. Initial research into categorical data analysis introduced two ideas for calculating multi-dimensional probabilities: the use of hash data structures as a means of storing and efficiently accessing LMI work breakdown structure (LMIWBS) combination data, and multi-way contingency tables (Dunham, 2003; Meyer, Hornik, & Zeileis, 2006). Exploration into the latter proved unsuccessful due to the vast number of possible combinations of LMIWBS entries, which number over seven thousand. Hash data structures enabled the retrieval of frequency counts for categorical data by storing records as a referenceable number, allowing rapid retrieval of that same data by calling upon only the reference number rather than the entire piece of information. These structures operate much like a card catalog system in the library; each book contains millions of characters but the computer system represents this data with only the aisle and shelf number, while the user still gets the entire text upon retrieval from the shelf. The speedy retrieval of frequency counts is a critical aspect of the computational approach needed for the categorical anomaly detection. Categorical data requires unique forms of analysis due to the fixed combinations possible to form entries rather than a more continuous form of information. To accommodate this requirement, we formed a likelihood measure (further described in Equation 3) rooted in conditional probabilities to determine if a record is anomalous.

Certain combinations of characters in the LMIWBS are problematic, which can lead to determining if a predicted action and object may be incorrect. An LMIWBS, such as RA102 (see Figure 1), which describes the assembly of non-aircraft fuel systems in a rotary wing platform, appears anomalous based on our identification process. Therefore, it has a higher probability of being an erroneous prediction than would an extremely common combination, such as RI021 (inspection of a structural component on a rotary wing platform) with over one hundred and fifty thousand occurrences. Simplistic approaches to anomaly detection consider only frequency counts, but these methods are not informative enough to show why an occurrence is potentially erroneous. The previous example, RA102, occurs only twelve times, which is relatively low given the size of the dataset. Conditional probabilities, however, quantify how likely it is for that specific combination to appear given a certain condition, such as *rotary wing*, signified by the lead character *R* in the combination. The LMIWBS is broken up into four components: end item, action, system, and subsystem; each unique instance of each category is represented by E_i, A_j, SY_k, and U_l, respectively. To provide the information needed to calculate conditional probabilities, a count function calculated the frequency of each unique component. Assessing the conditional probability of an event provides a deeper level of analysis, revealing not just that an occurrence is an anomaly, but also illuminating the specific segment of the combination that creates the anomaly. As an example, $P(E_i|A_j) = \frac{|E_iA_j|}{|A_j|}$ represents the conditional probability of the *i*th unique end item (E_i) existing in a maintenance record given that the *j*th action (A_j) exists in the same record. This probability is approximated by the frequency of the intersection of E_i and A_j ($|E_iA_j|$) divided by the frequency of A_j. Conditional probabilities found with two factors provide the first layer of insight, and a three-dimensional analysis will provide additional rationale for determining the cause of the anomalous behavior. For instance, looking at the number of occurrences of RA102, which is only twelve, does not provide much insight. However, using two- and three-dimensional conditional probability analysis, the process further reveals that both the probability of the action (assemble) given system (fuel system) and the probability of the system given subsystem (non-aircraft fuel system components) occurring together are under 0.02.

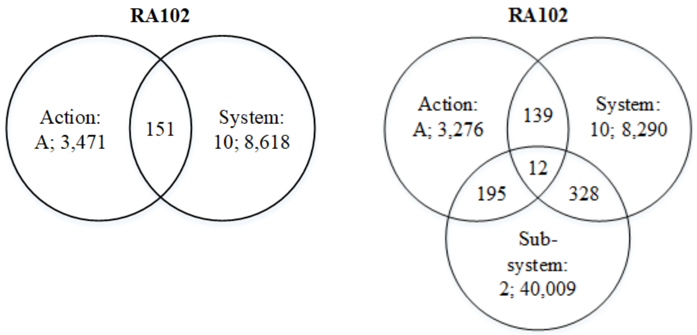

Figure 1. Example Venn Diagram Displaying the Unique Occurrences Elements Used for Conditional Probability Analysis

To further support this approach, we developed a likelihood measure. This method includes the conditional probabilities of all combination of LMIWBS elements. If the likelihood measure of an occurrence is low, relative to a free parameter determined by the organization performing the study, it is flagged as a potential anomaly for further analysis. In later steps, the algorithm will reserve flags for only those records that have a low probability of occurring on several different conditional probability measures. The following equation calculates the two-dimensional likelihood measure for an end item and action combination:

$$L(E_iA_j) = P(E_i|A_j) + P(A_j|E_i) = \frac{/E_iA_j/}{|E_i|} + \frac{|E_iA_j|}{|A_j|},\ given\ trials\ of\ E_i\ and\ of A_j > 30 \tag{3}$$

$L(E_iA_j)$ is compared to a free parameter, α, to determine if the combination is a rare occurrence with respect to the probability of events E_i and A_j individually occurring. For this study, we assumed an initial placeholder value for α of 0.1, which later proved reasonably effective in addressing stakeholder needs. However, we note that sensitivity analysis of the free parameter will prove crucial for more extensive applications of the model. Applying this concept to the previous example, the likelihood that RA102 occurs in terms of system and subsystem elements is 0.059, resulting in a flag if the organization conducting the study chooses an alpha value exceeding that measure. For this study, there are four dimensions, thus enabling the likelihood measure to capture up to three elements at a time. Since the study uses four elements, there are eleven elemental combinations that form the likelihood measures for each record in the dataset.

To further rank the results of the likelihood measures, we developed a set of "tests" for each LMIWBS. Each of the tests represent a unique arrangement of likelihood measures combinations. For example, the first test evaluates if the likelihood measures of both the action-system and action-subsystem are below the alpha value, as shown by the expression $L(A_jSY_k)$ *and* $L(A_jU_l) < \alpha$, and assigns a score of 1 if true. With each additional test failed, the LMIWBS combination's score increases by 1. Each unique LMIWBS is tested against an exhaustive list of likelihood measure combinations. For this study, there are only ten such tests with failures because no two-dimensional combinations featuring the end item scored below the alpha value. Although this solution will allow some erroneous records to pass through undetected and will provide many false positives, it allows analysts to target LMIWBS combinations that are unique in multiple categorical dimensions as they attempt to correct the data warehouse by hand or develop additional automated solutions.

3. Results

Results for both the scorecard model and anomaly detection process reveal that the MADW has been more accurate than the stakeholder initially projected. Most years represented in the scorecard analysis project a value of above 98, where 90 was originally expected. This expectation came from the only year in which the organization manually audited the accuracy of its data warehouse, fiscal year 2005, in which analysts determined that accuracy was at 90%. While the scorecard value is not a direct measure of accuracy by its technical definition, it is an aggregate of categorical values based on accuracy measures. Therefore, comparing the scorecard value to this accuracy estimate, while not directly correlated, still yields meaning for decision makers. Though this score has stayed relatively the same for recent years, the process of scoring an entire dataset provides a framework for analyzing data usefulness with respect to the stakeholder's needs. Note, however, that this method is limited to the thoroughness of automated data checks, which miss a good portion of the intuitively incorrect LMIWBS combinations. Our results for the scorecard model are summarized in Figure 2.

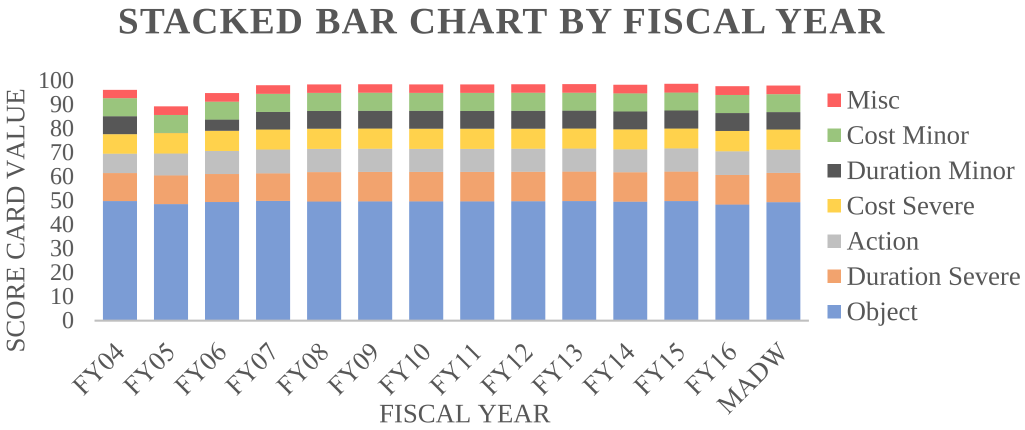

Figure 2. Stacked Bar Chart for the Fiscal Year Scorecard Value

With regard to anomaly detection, our method relying on conditional probabilities returned promising results. The process covered in this paper identified records that failed between one and ten tests in a sample dataset of rotary aircraft. This provided a means of determining which LMIWBS combinations are most likely to display erroneous records based on their anomalous observations (see Figure 3). To evaluate the performance of the predictive measure, we used a receiver operating characteristic curve (ROC curve), which judges the discriminatory ability of binary classification models (Hanley and McNeil, 1982). The area under the curve, 0.86 in this study, corresponds to the strength of classification, and represents the probability of a random negative instance being correctly ranked lower than a random positive instance. In order to validate the results, we analyzed each record with an LMIWBS that failed all ten tests, which suggests a high probability of error. After manually checking these records, we concluded that over 50% of those placed in the most anomalous group are in fact erroneous. In a dataset that features over 90% accuracy, finding such a high concentration of erroneous records is valuable as it prompts further investigation on the cause of erroneous entries with these specific LMIWBS combinations.

In an expanded trial, we randomly selected 8,000 entries from LMIWBS combinations failing four or more tests. From this segment of the study, we found a 62% errorneous classification rate, as compared to the roughly 4% error rate in the dataset as a whole.

Tests Failed	LMIWBS Combination Frequency
1	428
2	0
3	0
4	209
5	0
6	0
7	0
8	0
9	0
10	90

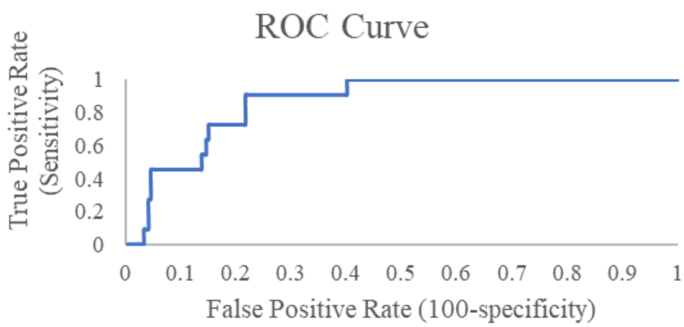

Figure 3. Summary of Test-Failure Results and the ROC Curve Plot

4. Conclusion

The presented research provides two main contributions: (1) a measurement of stakeholder-defined accuracy, and (2) a tool to detect errors in predictive processes within the categorical dataset. First, by looking at the historical accuracy of the MADW in accordance with automated checks, the score card model provides a way to assess the performance of a data warehouse. This allows the organization to identify areas of impact in which further effort will yield the greatest improvement in its system. Second, the concept of anomaly detection via analysis of conditional probabilities contributes to more efficiently identifying error-prone entries in a dataset. Currently, many organizations hand check their predictive measures to make corrections and then use those discoveries to narrowly modify the predictive algorithm on a case-by-case basis, using hundreds of work hours. The implementation of this method will narrow down the search area, reducing the hours spent searching for erroneous records. In addition to the immediate identification of these likely invalid records, this process may open doors to develop new automated solutions that can address erroneous records on the spot.

Moving forward, the predictive process can be improved by the inclusion of a system to identify common errors in anomalistic records, providing the groundwork to develop automation to self-correct these entries. In addition, the current model is based on our metric of tests-failed, which leaves room for the analysis of other measures. Further work exists in developing the ability to use not only failed tests, but also minimum, maximum, and average likelihood measures, or a combination thereof, to form a new predictive metric.

5. Acknowledgments

Parts of this paper were previously published and presented in the Donald R. Keith Memorial Capstone Conference at USMA in May of 2018. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Military Academy, United States Army, or United States Department of Defense.

6. References

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall/Pearson Education.

Barlow, H. B. (1989). Unsupervised Learning. In *Neural Computation*, volume 1, page 295-311.

Bayes, T., (1763). An Essay towards Solving a Problem in the Doctrine of Chances. In *Philosophical Transactions*. Vol. 53, page 370-418.

Bhaskaran, R., Palaniswamy, N., Rengaswamy, N. S., & Jayachandran, M. (2005). A review of differing approaches used to estimate the cost of corrosion (and their relevance in the development of modern corrosion prevention and control strategies). *Anti - Corrosion Methods and Materials, 52*(1), 29-41(13). Retrieved from https://search.proquest.com/docview/218922069?accountid=15138.

Das, K., & Schneider, J. (2007). Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 220-229). ACM.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Hanley, J.A., & McNeil, B. J. (1982). "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, *143*(1), 29-36.

Meyer, D., Hornik, K., Zeileis, A. (2006). The Strucplot Framework: Visualizing Multi-way Contingency tables with vcd. Retrieved February 22, 2018, from https://www.jstatsoft.org/.

Office of the Secretary of Defense. (2014). *Operating and Support Cost-Estimating Guide*. Cost Assessment and Program Evaluation. Retrieved September 25, 2017, from https://www.cape.osd.mil/files/os_guide_v9_march_2014.pdf.

Office of the Under Secretary of Defense (Comptroller). (2018). Operation and Maintenance Overview: Fiscal Year 2019 Budget Estimates. Retrieved October 27, 2018, fromhttps://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2019/fy2019_OM_Overview.pdf.

Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. (2010). *Prevention and Mitigation of Corrosion on DoD Military Equipment and Infrastructure*. Retrieved September 25, 2017, from http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500067p.pdf.

Shen, D., Ruvini, J. D., & Sarwar, B. (2012). Large-scale item categorization for e-commerce. Retrieved September 25, 2017, from https://www.researchgate.net/publication/262270957_Large-scale_item_categorization_for_e-commerce.

Yung, Chung. (2015). Mining Massive Web Log Data of an Official Tourism Web Site as a Step towards Big Data Analysis in Tourism. Retrieved September 25, 2017, from http://dl.acm.org/citation.cfm?id=2818906&CFID=985971970&CFTOKEN=42446460.