New Scoring Algorithm for predicting the cause of incident occurrences in chemical industries

Mr. Ganapathy Subramaniam B
Research Scholar, Software Development
CEI India Private Limited
Chennai, India

Dr. T. Ramaprabha
Professor, Department of Computer Science
VICAS
Tiruchengode, India

***Abstract*—Industrial incidents, which cannot be avoided in chemical and gas industries due to various factors involving operational errors, safety measures, defective workings etc. Occupational Health and Safety (OH & S) of India specifies the key regulations to make every place as an accident-free industrial environment. Recording and analyzing the historical industrial incidents data helps the industries to identify the potential trends and helps to reduce the loss. Cause of an industrial incident plays a major role to classify the incident to measure the severity and area of cautiousness. But in most of the industrial incidents, the cause of the incident occurred is unknown due to various factors. Modern Machine Learning techniqueshelp in this case to predict the cause of incident occurrences. This paper describes the new Scoring algorithm, very first version, to predict the cause of incident occurrences based on the various occurrence factors and impact factor.The algorithm has been trained and made learning with the help of FACTS incidents database which has the records of worldwide industrial incidents occurred between 2004 and 2014. With the help of Apache Spark, fast unified analytics engine and Microsoft .Net framework, the verification and validation of the algorithm has been done and the results of prediction have been discussed in this paper.**

Keywords—Workplace Incidents, Incident Occurrences; Scoring Algorithm, Predictive Model, Chemical industries incidents;Occupational Safety;

I. INTRODUCTION

Machine learning algorithms and its applications are ubiquitous in this modern world of Artificial Intelligence. The machinelearning processstarts with observations of data gathered from surveys, experience records, samples, to look for patterns in those to makethe decisions for the future. The aim of the machine learning algorithm is to make computers learn dynamically with very less or no human intervention and perform the actions accordingly. Machine learning helps the industries in the analysis of the regular flow of large quantities of data. While it has been deliveringquick and accurate results to identify positives and negatives, it may also require more time and area-specific materials to train them properly. Common machine learning algorithms do not work for all specific industries and organizations to produce the required results. Combining the ideologies of machine learning with cognitive technologies can make effective results by applying the customized algorithms in processing large volumes of industry data.

By including the various industry-specific factors into machine learning algorithms can provide advantageous impact for chemical and gas industries by reduced expenses, increased productivity, improved work methods. But industries have been slow to adopt the technologies available which might have to do with security concerns, cost, or even just a lack of understanding about the benefits to be gained.Evaluation of industrial incidental safety measures appears to be the weakest element of the industrial safety management system. Determination of the cause of the incident should help the industries by applying a precaution which in turn helps the supervision to provide the right solutions during the inspections. Formation of the new algorithm is an ideological tryto help the chemical and gas industries by including specific factors for determining the cause of the incident.The paper explains the algorithm, shows the analysis and implementation and concludes by providing the results of the same on applying in the incidents database which collected from chemical and gas industries.

II. LITERATURE REVIEW

The problem of classifying the unknown data into one of the segments is a fundamental one and has a wide, deep study to make it successful for many years.The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories [1].The introduction of Occupational Health and Safety (OH&S) management into the existing overall management system should be considered within a general management system model that incorporates the principles including the measurement and evaluation to take the preventive and corrective action [2].

An organization should measure, monitor and evaluate its OH&S performance, and take preventive and corrective action. Where appropriate, monitoring schemes for significant hazards should be in place. Broadly, such hazards may be classified as being either of the following: a) Physical, for example, noise, radiation, extremes of temperature; or b) Chemical, for example toxic, flammable or explosive [2].In establishing and maintaining procedures for investigating and correcting non-conformance, the organization should include these basic elements: a) Identifying the cause of the nonconformance; b) Identifying and implementing the necessary corrective action; c) Implementing or modifying controls necessary to avoid repetition of the nonconformance; and d) Recording any changes in written procedures resulting from the corrective action [2].The challenge which exists for the industries is that the complete relevance of the data. Based on the factors and occurrences, the columns which have unknown data or undetermined data should be predicted or calculated by means of the historical data learned by an algorithm.

III. METHODOLOGY

The generic form of any machine learning algorithm takes a function y(x) which generates an output vector y based on the input x. Training phase actually determines the definition of the function y(x) which is also known as the learning phase. Training has to be performed on the training data. Trained model can determine the desired result upon the data through the test set.Training data plays along well only with the less fraction of all input data. Customization of the training model in specific to the field helps to comprise the all possible input vectors to set the goal for achieving the predicted value in a long run.The sequence of study about incidents and inspections are necessary to build a machine learning algorithm which is an integral part to make the input data cleaner for prevention. It starts with the Inception phase where it has a high level sketch about the problem to resolve using the potential data inputs. Business knowledge helps to understand the key factors that are required for the problem.Using business knowledge, Data should be prepared, which might require some complex transformations as well, for a model to observe and learn from them. The goal of the model is to achieve the primary objective by measuring the values to predict the cause of the incident. It needs enormous work to transform the raw data into potential model inputs by performing cleaning up the bad data, removing outliers, segregation of the quality data. Once data is available, the prototype model should be built for experimenting with the solution.

Underlying calculations of the algorithm to provide the expected results is not a straightforward equation. By continuous iteration process of validation, trail & error and improvisation, the algorithm can fit into a production-ready model. Assessing the algorithm's performance based on the predefined quality factors would be done by comparing the results of prediction. Selecting the most relevant features in turn to obtain the impact factor would affect the performance of the algorithm.Building the model involves time-consuming tasks like labeling the data, verification,and validation of the model. Labeling a huge amount of data points with relevant categories as input for a classification algorithm to perform testing the output of the model is correct. These tasks should be done on the fly as they come up by building the data pipelines and streaming apps.

IV. ALGORITHM

Given an unknown function $y = f(x) + e$, where f is the learning target function that has input variables *x*which is an occurrence in this model, to produce the desired result in *y*which is the cause of the

incident. *e* is an error that is independent of the input variable *x*.Output function maps input instances $x \in X$, a collection of the historical incident datato output $y \in Y$, a known cause of the incidents, along with training data $\mathcal{D} = \{(x_0, y_0),\ldots(x_n, y_n)\}$. With these input elements in hand, the primary goal is to derive an algorithm to predict the cause of the incident with reference to the facts of incidents reports and inspection data about the occurrences. Calculating the score value based on the historical occurrences can be achieved using the following formulae:

$$S_{(cx)_{ij}} = \frac{1}{t} \bullet \frac{\sum_{i=1}^{n} cx_i}{\sum_{j=1}^{n} xc_j} \bullet x_{ij} \ldots\ldots\ldots\ldots (1)$$

where *S* is a score calculated for each occurrence (*x*) for each cause (*c*) and x_i is a number of occurrences accounted in the incident and *t*are a number of iterations performed as shown in the formula*(1)*. Following is the formula to obtain the predictable value of cause of the incident.

$$y_i = c \left[\sum_{i=1}^{n} Scx_i \bullet O_i\right] \ldots\ldots\ldots\ldots (2)$$

where *O* is the given occurrence for the prediction and *S* is the scoring value obtaining from the training data, c is the corresponding potential values of cause and *x* is the historical occurrence summation value. Pseudocode of the algorithm is as follows:

```
algorithm scoring-cause is
input: Occurrences O with inspections data i,
        impact constant k,
        node t
output: Cause c such that c is maximal from 0 to sum of t
for eachfactoroccurs(f,o) inalldatado
      s(c,o) ← 0
      sum_score ← 0
      inter_score ← 0
      for eachcauseoccurs(c,o) in alldatado
      inter_score ← Σ v(c,o)
      sum_score ← v(f,o) / inter_score
      s(c,o) ← (v(f,o) / k) + sum_score * 1 / t
maxf ← 0
for eachpd(f, o)ininputdatado
      if maxf< (Σv(f,o) * s(c,o))
      maxf ← (Σv(f,o) * s(c,o))
        c = cause(maxf)
return c
```

V. RESULTS

Data for training and testing are collected from factsonline.nl, FACTS (**F**ailure and **Ac**cidents **T**echnical information **s**ystem) is an incidents databases which have 25,700 chemical industrial incidents which provide the level 0 access to free search and insight of the possibilities in basic accident descriptions with the years span from 2004 to 2014. A similar subset of the dataset is also available in data.gov.in which is an open government data platform of India. Data, obviously, has got unknown-cause of the incidents which gives us the way to predict the cause by utilizing the data which has known-cause as training data. Implementation of the algorithm is done using C# Windows application and experimented the prediction of the cause.

Implementation of the algorithm is done in two phases. First one is to generate a score followed by predicting the cause of the incident. Figure 1 shows the initial screen of the implementation application with the steps to follow: 1. Training the support vector, 2. Classify and predict and 3. Do it in the loop for all the incoming data. The flow of data through the data frames have been done using Kafka and Apache Spark.

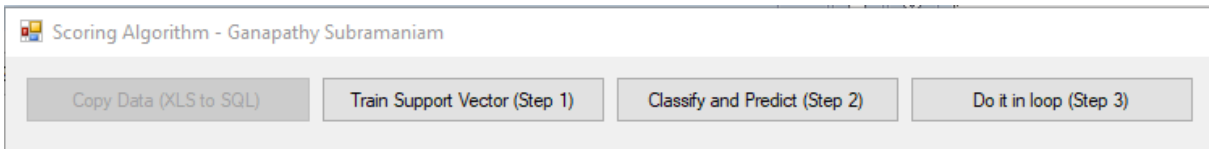

Figure 1: Initial screen

Scores obtained from the first step of the Scoring algorithm is shown below in Figure 2.

	Cause	Total	Factor	Dust_emission	Chemical reaction	Overtum_Capsize	Fill	Dry	Impl
▶	Domino	5	1	0	0	0	0	0	0
	Human-failure	6331	4	0	0.292875636103...	0.687027227608...	0.365361657109...	0.250315905860...	0.14
	Management-fail...	7919	1.2	0.728282955836...	0.366876581913...	0.078575481051...	0.364379751965...	0.250252557141...	0
	Natural-cause	1505	10	0	0.045110501227...	0.098084534804...	0.024584717607...	0	0.28
	Op/Others-ext/Fail	129	1	0	0.004468497017...	0.009810719851...	0	0	0
	Sabotage/Vanda...	624	1	0	0.013405491051...	0.024526799627...	0	0	0
	Technical-failure	7529	1	0.273089285518...	0.312794791213...	0.142255437842...	0.258468713956...	0.500482683721...	0.57

Figure 2: Score data

Predicting the cause of the incident using the given input data which utilizes the full algorithm is shown in Figure 3.

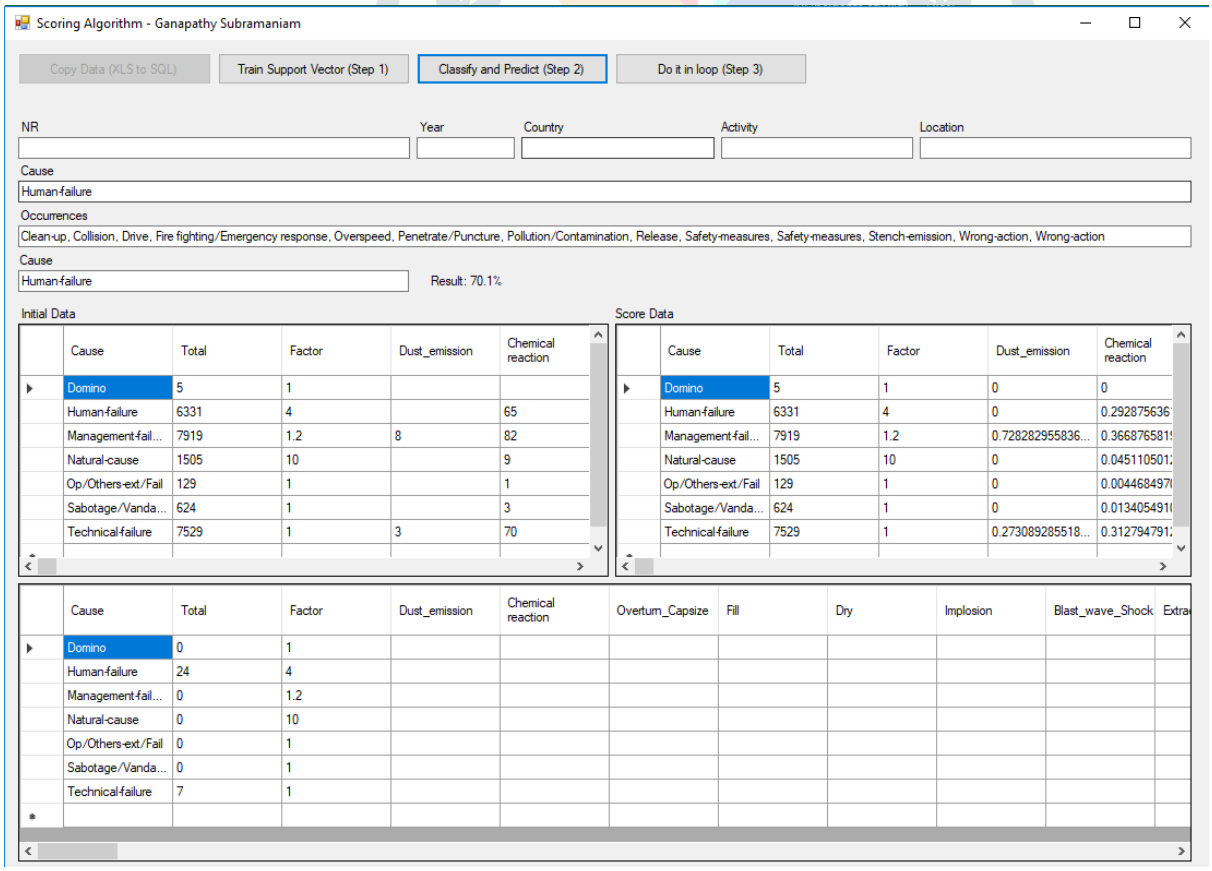

Figure 3: Complete implementation

The algorithm shows the promising results with the given data under test and overall result based on the initial iteration training data spans between 67.3% and 70.1%. Further tweaks in the learning mechanism and factors would improve the results where the algorithm has full potential to predict the cause by 95% or closer.

VI. Conclusion

Customized Support vector algorithms can bring-in the greater advantage in chemical and gas industries to provide the cleaned up data for further processing like deep analysis and prediction. The objective of the algorithm is to bring an unbiased, close-enough see through the data to obtain a logical conclusion to find the cause of the incident. The methodology may not precisely explain about all variations of the technique but gives a quick walkthrough of the algorithm and its objective. Further study can be done on the model to compare with other models to obtain a better insight into the prediction of the cause of the incidents. Continuous observation is also being performed and closely monitored forthe performance of the algorithm.

VII. References

[1]. Bishop, C. M. (2013). *Pattern Recognition and Machine Learning*. *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1117/1.2819119

[2]. Bureau of Indian Standards. (2007). *Occupational Health and Safety Management Systems*, *2000*, 1–28. https://doi.org/10.1002/0471435139.hyg049.pub2

[3]. Dietterich, T. G. (2009). *Machine learning in ecosystem informatics and sustainability*. *IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.1007/978-3-540-75488-6_2

[4]. Janicak Christopher. A. (1996). *Predicting accidents at work*, 115–121.

[5]. M Surianarayanan and G Swaminathan (2000). *Indian chemical industry accident database – an effort by CISRA Cell for Industrial Safety and Risk Analysis Chemical Engineering Department*, Chennai.

[6]. Mitchell, T. M. (Tom M. (1997). *Machine Learning*. McGraw Hill.

[7]. Safety, C. S., & Hazards, C. S. (2017). FACT SHEET » Confined Spaces.

[8]. Yael Gavish. (2017). *Developing a Machine Learning Model from Start to Finish*.

Ganapathy Subramaniam B., from Chennai and born in 1979, research scholar from Periyar University in the field of Computer Science since Feb 2017. Data mining is his major field of research. He has obtained a Masters of Philosophy in Computer Science from Alagappa University in the year of 2006 and completed his Master of Science in Information Technology in 2002. He has earned a bachelor's degree in Commerce from Annamalai University in 2000 and has also obtained a Diploma in Computer Technology from All India Council of Technical Education in 1997. His field of study is data mining in gas detection domain.

Dr. T.Ramaprabha, M.Sc.,M.Phil.,Ph.D., works as a Professor of PG and Research Department of Computer Science and Applications in Vivekanandha College of Arts & Sciences for Women(Autonomous), Tiruchengode, Tamil Nadu. She has completed her Ph.D., in the year 2013 at Mother Therasa Women's University, Kodaikanal. She has more than 22 years of teaching and research experience and she has published many papers in conferences and journals. Her researchof interest is stereo pair image compression in image processing.