Self-Augmenting Knowledge Base for Informed Decision Making with biomedical applications in cancer diagnosis

LES SZTANDERA
Kanbar College of Design, Engineering, and Commerce
Thomas Jefferson University
Philadelphia, PA 19144
USA

Abstract: - Fuzzy sets methodology to automatically generate knowledge base for informed decision making is proposed. As a proof of concept it has initially been applied to generate regulatory/health/environmental guidance rules for textile and apparel companies. Subsequently, the system will be augmented to incorporate additional consumer goods, and down the road, after some modifications, could be utilized as a much needed health care disruptor tool in personalized medicine for both patients and clinicians. The apparel category provides for a diverse set of mandatory regulations and some voluntary standards. Mandatory requirements such as CPSIA, FTC for Care and Textile labelling, in addition to AATCC requirements for colourfastness and formaldehyde were taken into consideration. Initial focus was on carcinogenic dyes and pigments. Databases from the International Agency for Research on Cancer (IARC), the US National Toxicology Program (NTP) are to be incorporated, in conjunction with computational intelligence, to identify potential toxins or carcinogens present in the industrial process or the final product, thus alerting manufactures and consumers through a user-friendly interface. This capability can be quickly developed and validated using modern software product development approaches incorporating Design Thinking, Agile Development with Scrum, and Business Model Generation to get this to market where key benefits can be derived.

Key-Words: - Fuzzy Sets, Knowledge Base, Decision Making, Biomedical Applications, Cancer Diagnosis

Received: November 26, 2019. Revised: February 28, 2020. Accepted: March 30, 2020. Published: April 10, 2020.

1 Introduction

For most real-world decision-making systems and problems behind them, the information concerning the design comes either from an experienced human user or from crisp sensor measurements. Suppose, the environment facing the decision maker is so complex that no mathematical or statistical model exists for it. The task here is to design a fuzzy rule generation system to replace the user development of the system. The proposed system will advise companies (initially textile/apparel businesses to establish proof of concept) regarding the multitude of mandatory and voluntary regulations (*e.g.* H.R.2576, CPSIA, FTC for Care and Textile labelling, AATCC requirements for colorfastness, and formaldehyde usage). In addition to regulatory advice, the system is unique in that it will identify potential hazardous chemicals/materials present during the manufacturing process or linked to the final product. This will be accomplished by coupling the system to databases from the International Agency for Research on Cancer, the US National Toxicology Program, and to a computational chemistry/intelligence CODESA-based protocol for identifying toxins / carcinogens / allergens / environmental hazards. Our system will serve as an automated tool to efficiently and accurately identify global regulatory/testing/safety requirements and alert consumer-product manufacturers of potential problems at every stage of the development cycle. This innovative service will assist companies (initially, textile/apparel businesses) avoid costly ligation by retrieving all relevant documentation related to the products they produce. Uniquely, our system will also screen the materials/chemicals involved at various phases of production for possible public health/environmental concerns, reporting any warnings connected to related materials in a coherent and cohesive way to the user. This

inclusive tool will facilitate manufacturers across the industrial spectrum create more hazard-free products for consumers. It will serve as a much-needed tool in preventive health care. The final development phase will incorporate many additional features including data about tumour tissue DNA, enabling the system to serve as a much-needed health care disruptor tool in personalized medicine for both patients and clinicians.

2 Problem Statement

There exists an industry need to further develop a decision-making solution to provide an automated process for the identification of global product safety requirements based on pre-defined product criteria with a tool to enhance the efficiency and accuracy of determining reasonable testing programs to insure safety of products being placed in the global market place. The escalation and complexity of regulations and testing requirements, as well as their access is of vital concern to decision makers in the textile/apparel industries. Currently the process of identifying required testing regulations is accomplished by an experience user and is done by hand, making this process time consuming and prone to oversights. In addition, the current set of recommendations does not cross-reference International Agency for Research on Cancer (IARC) and the US National Toxicology Program (NTP) databases.

3 Proposed Solution

The unique advantage of the proposed technology lies in harvesting hidden relationships through the computational power of fuzzy sets. The proposed tool provides the end user with unique knowledge to satisfy regulatory requirements on one hand, and to differentiate and add value to the consumer products on the other hand. The driving force behind this differentiation lies in utilization of fuzzy sets. This approach is much more adequate for dealing with uncertainty and complexity of factors involved.

The scope of this research is to create a prototype knowledge base system to integrate industry regulations as well as to address consumer concerns, expressing them as fuzzy rules, and subsequently translating and displaying them through a friendly user interface. A software package that integrates and summarizes required testing/regulatory information, ATTCC requirements, data on potential health/environmental concerns, as well as recommending improved, cost-effective materials/processing procedures, and presents the results in a coherent and cohesive way to the user, all in one place, will be invaluable to textile/apparel companies. We have worked closely with industry to assure the relevance, accessibility, and potential impact to a large segment of the textile/apparel industries. Incorporating the identification of toxic agents in consumer textiles/apparel and their processing makes this knowledge based system unique [1]. The innovation lies in integrating computational intelligence and computational chemistry to identify potential toxins and carcinogenic agents and alert consumers/manufacturers. These approaches are much more robust in dealing with uncertainty and complexity of factors involved in potential toxin formations than any statistical analysis and tools currently utilized in chemical compound recognition and classification.

3.1 Methodology

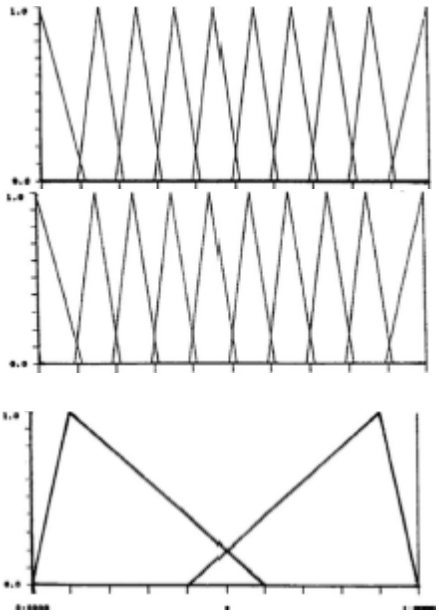
The method consists of the following four steps: Step 1) divides the input and output spaces into several fuzzy subsets and assigns linguistic terms to them; Step 2) generates fuzzy rules using the linguistic terms assigned in Step 1; Step 3) counts the conflicting fuzzy rules and those with the highest number of counts remain in the system; others are deleted; Step 4) determines a mapping based on the remaining rules. Suppose the data set under investigation consists of the multi- dimensional data where the first coordinates form the input, and the last one is the output. The goal is to design a classifier capable of predicting the output given new multi-dimensional inputs. So, we have several input variables and one output variable.

We must know the ranges for both input and output.

Our task is to generate a set of fuzzy rules from the training input-output rows and use these fuzzy rules to model the system. We proceed in the following manner. First, we divide the input and output spaces into several fuzzy subsets and assign linguistic terms to them. Let's assume that the domains of input and output variables are provided by a feeder computational chemistry approach. Utilizing it, Quantitative Structure-Activity Relationships (QSARs) are developed that correlate the observed mutagenic activity of 181 aromatic amine derivatives with a variety of molecular descriptors calculated using quantum-chemical semi-empirical methodology. Conventional multiple linear regression techniques using five descriptors give a relationship that accounts for approximately 66% of the observed variation in the relative mutagenic behavior of these compounds; increasing the number of descriptors does not significantly improve the correlation equations. Our approach using fuzzy sets can account for more than 90% of this variation using ten descriptors [2, 3]. The generated fuzzy rules will subsequently be augmented by multiple existing databases to form a knowledge base for informed decision making. In the example above ten molecular descriptors calculated using quantum-chemical semi-empirical methodology are divided into ten regions each and are assigned the following linguistic terms: very very low, very low, low, low-medium, medium, medium-high, high, very high, very very high, and extremely high. The domain of output variable, mutagenic activity, is divided into two regions with linguistic terms low and high assigned to them. Next, we assign to each region a fuzzy membership function. Different shapes of membership functions are possible, but we use triangular shapes with height 1 at the center of the region, and 10% overlap between neighboring sets. The next step is to generate fuzzy rules using the linguistic terms assigned in the previous step.

Since the antecedents are different components of a single input vector, the rules are in the form of "IF ... AND ... 1HEN ...", where the "IF ... AND ... " part is generated from the input data, and the 'THEN" part is generated from the output data. As the data may be conflicting, and so far we have generated one rule for each data row, the following heuristics is employed. The conflicting fuzzy rules are counted and those with the highest number of counts remain in the system; others are deleted; however, some could be re-inserted into a system allowing for self-augmentation. Fig. 1 shows the membership functions used along with sample rules. In the last step, fuzzy inference engine is executed and a mapping based on the remaining rules is determined. Fuzzy inference is a method that interprets the values in the input vector and, based on some sets of rules, assigns values to the output vector. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made. Two main types of fuzzy inference systems can be implemented: Mamdani-type and Sugeno-type [4]. These two types of inference systems vary somewhat in the way outputs are determined. Mamdani-type inference expects the out membership functions to be fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable, which needs de-fuzzification. De-fuzzification is the process of obtaining a single number from the output of the aggregated fuzzy set. It is used to transfer fuzzy inference results into a crisp output. De-fuzzification is realized by a decision-making algorithm that selects the best crisp value based on a fuzzy set. There are several forms of de-fuzzification, including center of gravity (COG), mean of maximum (MOM), and center average methods. The COG method returns the value of the center of area under the curve and the MOM approach can be regarded as the point where balance is obtained on a curve.

In our case de-fuzzification is based on the center of gravity method which is sensitive to all the remaining rules [4, 5, 6].



Sample Rules:
IF x is very very high AND y is extremely high THEN z is high
IF x is very very low AND y is extremely high THEN z is high
IF x is medium AND y is medium-high THEN z is high

Fig. 1. Sample membership functions along with "IF ... AND ... THEN ... " rules.

3.2 Knowledge Base Development

As the proposed approach calls for utilization of both empirical data and multiple existing databases, the plan is to use a Lean Start-up [7] approach to learn as quickly as possible. This method focuses on the use of Design Thinking methods, Agile Software Development (Scrum) [8], and creating a Minimum Viable Product (MVP) to validate the solution with users. The process will begin by developing the product strategy with input from Design Thinking techniques. Interviews with end-users, textile manufacturers, regulatory advisors to fully understand and capture the problem to be solved, and an idealized solution from those closest to it. The next step would be to develop a product strategy using Business Model Generation [11] to capture key elements that will help scope the MVP and determine the business model to support its development. These elements will include Product Vision, Target Customers/Users, Needs, Business Goals, Competition Map, Regulatory Requirements, Funding Sources, Cost Factors, and Channels. This will provide a validation framework for strategic decisions. With that information in place, we can start building user journeys and experience maps from the perspective of various stakeholders. This will result in a product roadmap containing prioritized features that can then be estimated and a plan established for the initial MVP. This will feed the agile product development process. With the above components in place the development team can start creating the highest value features and attacking the highest risk items. In taking this approach, it allows for an iterative and incremental process where key features are put in front of users as quickly as possible to get feedback early and often. This is accomplished by using time-boxed iterations of 2-3 weeks which include user and stakeholder product review sessions. The benefits are increased engagement in solving the problem, and ensuring that changes from a fast-moving complex regulatory/economic environment are brought into the picture as early as possible. The approach also works as a continuous validation method as features are fully tested in the iterations avoiding the build-up of technical debt that is common in early prototypes. The iterations will continue until the team agrees that an MVP is in place that can operate in a production environment where derived benefits can be fully validated from the original goals and the value proposition can be solidified. This approach will answer questions about what are the full complement of data sources, how that data can be combined with computational intelligence, chemistry QSAR calculations, and other data to create a knowledge base that will provide users with faster, more accurate testing requirements that identify potential safety problems much more efficiently.

The key resources will be interested parties from textile manufacturing, textile/apparel testing companies, regulatory expertise (country specific), cloud-based hosting services, development environment, and software development services.

The first year will be comprised of two major phases:

Phase 1: 4-6 months (2-week iterations)
Interviews of Users and Stakeholders to create User Experience Journeys
Identify Key Risk areas for exploration
Create product strategy using Business Model Generation
Create Product Roadmap
Get Team in Place for Phase 2

Phase 2: 6-8 months (2-week iterations)
Establish Technical Infrastructure and Cloud Hosting Environment
Begin Development of Features
Feedback at End of Each Iteration
Revise roadmap adjusting for feedback
Work to establish MVP as Proof-of-Concept
Develop Product Commercialization Strategic Plan

The use of the Lean Start-up method is a proven approach to conduct research to determine the viability of a problem solution in the market as quickly as possible with the minimal amount of investment. By first developing the product strategy with engaged stakeholders and affected users, we can identify much earlier the minimum product features that will provide the most value. The research will test the assumptions about available regulatory data, computational intelligence calculations, and the regulatory process itself to determine if a knowledge base can provide anticipated benefits in both speed and accuracy. And if that data can be intelligently combined and presented to users in a simple, intuitive way, we can measure the benefits to manufacturers, regulatory bodies, and increased safety to consumers. After the first year, if the research and product MVP are validated, and a business model can be established, the product could be generally available in 3-6 months (total of 10-14 months). Development would continue to build out additional features and incorporate increased feedback. The next phases would begin to explore additional applications in expanded industries such as healthcare. See also [9] and [10].

4 Conclusion

In this paper, we proposed a general method to generate membership functions and "IF ... AND ... THEN ..." rules from empirical data calculations. The main features and advantages of the generation method are: 1) it is a simple and straightforward quick-pass build-up procedure; hence, no time-consuming iterative training is required; 2) there is a lot of freedom in choosing the membership functions; this provides us with a lot of flexibility to design systems according to different requirements. We believe that the system designed by using our method can perform successfully for cases where numerical data from either empirical calculations and/or existing databases are provided.

The unique advantage of the proposed technology lies in harvesting hidden relationships through the computational intelligence power of fuzzy sets. The proposed tool provides the end user with unique knowledge to satisfy regulatory requirements on one hand, and to differentiate and add value to the consumer products on the other hand. The driving force behind this differentiation lies in utilization of fuzzy sets approach. This approach is much more adequate for dealing with uncertainty and complexity of factors involved. Using the proposed knowledge base will enable decision makers to retrieve all relevant documents related to regulatory/testing information, AATCC requirements, data on potential health/environmental concerns, as well as provide suggestions on improving materials and development practices. The system will present its findings in a coherent and cohesive way to the user, thus enabling informed decision making. The combination of an innovative method to solve this complex problem space; along with modern software

product development techniques will make sure that the right product is available as quickly as possible to address the critical need in the apparel manufacturing industry. This inclusive tool will facilitate manufacturers across the industrial spectrum create more hazard-free products for consumers.

5 Future Research

This application could also serve as a much-needed tool in preventive health care. The final development phase will incorporate many additional features including data about tumour tissue DNA, enabling the system to serve as a much-needed health care disruptor tool in personalized medicine for both patients and clinicians. We envision that the proposed knowledge based system could provide an indispensable tool for clinicians to identify potential cancer type in patients who have a certain genetic makeup by running corresponding compound matches. The proposed knowledge base requires using existing fuzzy rules generating software to provide entries for the system.

References:

[1] L. M. Sztandera and C. W. Bock, Intelligent Quality Compliance System for Textile and Apparel Businesses, *International Conference on Innovative Engineering Technologies (ICIET),* 1, Krakow, Poland, January, 2017.

[2] K. L. Bhat, S. Hayik, L. M. Sztandera, and C. W. Bock, Mutagenicity of Aromatic and Heteroaromatic Amines and Related Compounds: a QSAR Investigation, *QSAR Combinatorial Science,* 24(7), 831-843, 2005.

[3] L. M. Sztandera, A. Garg, S. Hayik, K. L. Bhat, and C. W. Bock, Mutagenicity of Aminoazo Dyes and their Reductive-Cleavage Metabolites: a QSAR/QPAR Investigation, *Dyes and Pigments*, 59(2), 117-133, 2003.

[4] L. M. Sztandera, Extracting information from failure equipment notifications - use of fuzzy sets to determine optimal inventory, *WSEAS Transactions on Computers*, 10(11), 390-395, 2011.

[5] J. Sanford and L. M. Sztandera (2011), Classification data mining with hybrid fuzzy logic aggregation, *WSEAS Transactions on Computers*, 10(11), 379-389, 2011.

[6] Wang L. X. and Mendel J., Generating Fuzzy Rules by Learning from Examples, *IEEE Transactions on Systems, Man, and Cybernetics,* 22(6), 1414-1427, 1992.

[7] Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Business.

[8] Schwaber, Ken and Jeff Sutherland. *Software in 30 days: how agile managers beat the odds, delight their customers, and leave competitors in the dust*. John Wiley & Sons, 2012.

[9] Anna Walaszek-Babiszewska, Application of a Stochastic-Fuzzy Approach to Modeling Optimal Discrete Time Dynamical Systems by Using Large Scale Data Processing, *WSEAS Transactions on Systems,* Volume 18, 2019, pp. 144-148

[10] Teimuraz Tsabadze, Archil Prangishvili, One Approach to using Fuzzy Logic for the Establishing of Natural Gas Tariffs, *WSEAS Transactions on Systems,* Volume 18, 2019, pp. 138-143

[11] Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: a handbook for visionaries, game changers, and challengers*. John Wiley & Sons.