Analysing Food Allergy Dataset to Predict Different Types of Allergies

Mrs. B. Sowmiya¹, Rahul Ramaswamy², Tanmai Shukla³

¹²³AP/CSE, SRMIST, Chennai, India
Department of CSE, SRMIST, Chennai, India

Abstract- To analyse allergy dataset and find which food item contains allergy causing chemicals. Nourishment hypersensitivities have dependably been a general medical issue yet are winding up more applicable around the world. There are a few chemicals which are present in different food items causing a particular kind of allergy. The main idea of the paper is to analyse those food items and put them into different groups based on the allergy being caused, which can help people choose which food item to avoid easily. With no fix accessible, individuals experiencing food allergies need to entirely maintain a strategic distance from real allergens, for example, eggs, shellfish, nuts, wheat, and soybeans. That is on the grounds that even somewhat undesirable or unfavorably susceptible substance can cause serious indications in a few people, possibly prompting hypersensitivity and death. The paper aims to survey on the above mentioned problems.

Keywords- Clustering, Food allergy, K-medoids, Data Mining, Naïve Bayes

1. Introduction

As a recently perceived illness, sustenance hypersensitivity is causing increasingly consideration among general society and around the globe. Individuals probably won't have seen, yet nourishment sensitivity has turned out to be more pervasive and discredit its significance is dangerous. Sustenance sensitivity exists together with asthma, atopic dermatitis and can cause hypersensitivity response. Contrasted with grown-ups, nourishment sensitivity could be more hazardous to kids, in light of the fact that the hazard and chance that youngsters inadvertently eat their sustenance allergens are higher. Proof demonstrates that hospitalizations for hypersensitivity have expanded a ton among youngsters, with sustenance initiated hypersensitivity as the most widely recognized reason.

Notwithstanding the seriousness and expanding pervasiveness of nourishment hypersensitivities, there is no present treatment to fix sustenance sensitivity completely. Numerous individuals have not yet perceived what nourishment sensitivity truly is and how genuine it could be. Additionally, there are misinterpretations about nourishment narrow-mindedness and sustenance sensitivity because of the absence of right information. Subsequently, it is first critical to perceive that sustenance hypersensitivity is an ailment that could prompt serious responses, and to comprehend what may cause it and why. At that point finding out about the present medications which desensitize nourishment hypersensitivities and the right responses to other people who get into hypersensitivity is fundamental. There are additional things that the general population and everybody can do to assist individuals with sustenance hypersensitivities.

All the papers use a mixture of many datasets which are not related to each other well. A single large dataset is nowhere used which will yield the best results. The source of datasets are also different which includes twitter data, clinic data, etc.

The algorithms mainly used by the researchers are K-medoids algorithm, which is a clustering algorithm similar to the k-means algorithm. Naïve Bayes classifier which belongs to the group of simple "probabilistic classifiers" in view of applying Bayes' hypothesis with solid autonomy suspicions between the highlights, the k-closest neighbors calculation which is a non-parametric technique utilized for characterization and worsening, Artificial Neural Networks (ANN) which is one of the commonly applied machine learning algorithm and SVM are directed learning models with related learning
calculations that break down information utilized for arrangement and relapse examination, all of which are suitable for the given application.

2. Related Works

The following are the different approaches used by the developers and researchers for predicting various allergy causing chemicals.

A. “Using Mutual Information Clustering to Discover Food Allergen Cross-Reactivity”

“Kenneth H. Lai, Suzanne V. Blackley” et al (2017) have used K-Medoids and mutual information clustering to club random variables or sets. A prologue to another standardization parameter to control the extent of the groups, and applying it to nourishment sensitivity information from a vast hypersensitivity vault has been finished [1].

B. “Public Health Allergy Surveillance Using Micro-blogs”

“Kruti Nargund, Dr. Natarajan S.” (2016) have used Naïve Bayes Classifier, Support Vector Machine and KNN classifier to analyse wellbeing related tweets and the primary spotlight is on sensitivity illness. The paper also concludes that KNN is better than other classifiers. [2]

C. “The soft computing-based approach to investigate allergic diseases: a systematic review.”

“Gennaro Tartarisco, Alessandro Tonacci”, et al (2017). The main aim of the paper is to symmetrically audit the principle delicate processing based systems, for example, ANN, SVM, and so forth to examine their execution in the field of unfavorably susceptible illness. [5]

D. “A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin test”

J. “Jabez Christopher, H. Khanna Nehemia” et al (2015) use “Clinical Decision Support System” with rule based grouping way to deal with help specialists in the finding of hypersensitive rhinitis to settle on solid choices, based on intra dermal skin test. [6]

3. Algorithms Used

1. The k-medoids calculation is a bunching calculation like the k-implies calculation and the medoid move calculation. Both the k-means and k-medoids calculations are partitional and both endeavor to limit the separation between focuses named to be in a group and a point assigned as the focal point of that bunch. Its advantages are that it can solve K- means problems and generate empty clusters and is sensitive to outliers and/or noise. It also selects the most centred member belonging to the cluster. Its disadvantages are that it requires precision and is quite complex.

![Cluster plot]

Figure 3.1
2. Naïve Bayes - In machine learning, Naïve Bayes classifiers are a group of straightforward "probabilistic classifiers" in light of applying Bayes' hypothesis with solid autonomy suppositions between the highlights. Its advantages are-

- Very simple, easy to implement and fast.
- Need less training data.
- Handles continuous and discrete data.

3. SVM- In the field of machine learning, support vector machines are managed learning strategies with related learning calculations that dissect information utilized for characterization and relapse investigation.

SVMs are very good when the data is unknown.

- Works well with unstructured and semi structured data like text, Images and trees.

![Figure 3.2](image)

4. KNN- In pattern recognition, the k-closest neighbors calculation is a non-parametric strategy utilized for grouping and relapse. In both the cases, the information comprises of the k nearest preparing models in the element space.

![Figure 3.3](image)

5. ANN- Artificial neural networks or connectionist frameworks are figuring frameworks dubiously roused by the natural neural systems that comprise creature minds.
ANNs can learn and demonstrate non-direct and complex connections, or, in other words on the grounds that, all things considered, a significant number of the connections among information sources and yields are non-straight and also unpredictable.

6. Fuzzy Logical model - Fuzzy logic is a type of many-esteemed rationale in which reality estimations of factors might be any genuine number somewhere in the range of 0 and 1. It is utilized to deal with the idea of fractional truth, where reality esteem may run between totally obvious and totally false.

4. Analysis Report

Table 4.1
| S.No | Publication Year | Author Name | Paper Title | Algorithm Used | Dataset | Accuracy |
|------|------------------|-------------|-------------|----------------|---------|----------|
| 1    | 2017             | “Kenneth H. Lai, Suzanne V. Blackley, Li Zhou” | “Using Mutual Information Clustering to Discover Food Allergen Cross Reactivity” | Mutual Information Clustering, K-Medoids | EHR Data Repository | Max ARI value of 0.70 2 |
| 2    | 2016             | “Kruti Nargund, Dr. Natarajan S.” | “Public Health Allergy Surveillance Using Micro-blogs” | Naïve Bayes Classifier, SVM, KNN classifier | Twitter Dataset | KNN precision value of 0.86 4 |
| 3    | 2016             | “A. Velmurugan, T. Ravi” | “Allergy Information Ontology for Enlightening People” | Precision, Recall and F-measure metrics | Allergy Clinic Dataset | P, R and F value of 0.85 |
| 4    | 2014             | “Ha X. Dang, Christopher B. Lawrence” | “Allerdictor : fast allergen prediction using text classification techniques” | SVM based allerdictor | International Union of Immunological societies allergen nomenclature, Allergome, SDAP, allergen online, Aller match | AUPRC value of 0.85 5 |
| 5    | 2017             | “Gennaro Tartasco, Alessandro Tonacci, Paola Lucia a Minciullo, Lucia Billeci” | “The soft computing-based approach to investigate allergic diseases: a systematic” | Bayesian Network, Fuzzy Logic model, SVM, ANN | ScienceMed PubMed | Accuracy of 86.5% |
| Page | Year | Authors | Title | Methodology | Acronym | Results |
|------|------|---------|-------|-------------|---------|---------|
| 6    | 2015 | J. Jabez Christophe r, H. Khanna Nehemia, A Kannan | “A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin test” | CDSS, ANN | Allergy Testing centre | Accuracy=88.3% Sensitivity=88.3% Specificity=88.2% |
| 7    | 2013 | Raquel Gutiérrez Rivas, Juan Jesús García Domínguez, William P. Marnane, Niall Twomey, Andrey Temko | “Real time allergy detection” | QRS Detector Algorithm | Clinical Research Ethics Committee of the Cork Teaching Hospital | Specificity=100% Sensitivity=100% |
| 8    | 2015 | Kathy Lee, Ankit Agrawal, Alok Chaudhary | “Mining Social Media Streams to Improve Public Health Allergy Surveillance” | Bag-of-Words, Naïve Bayes multinomial classifier, Spatiotemporal Analysis | Twitter dataset, Pollen Dataset, Climate Dataset, Allergy Dataset | Precision of 86.7% |
The given table 4.1 is the literature survey showing the work of various researchers in the field of allergy prediction using various algorithms. It mainly compares a few main algorithms to check for their accuracies and precision. Some researchers have also modified the existing algorithms as per the needs of their work.

Some of the work done above are on a specific allergy, whereas some have worked on generalised datasets. The main objective of all the above researchers is to find the various allergy causing chemicals and let the people know what the main cause of allergy is and how to avoid them.

5. Conclusion and Future Enhancement

The “U.S. Food and Drug Administration (FDA)”, reports that a huge number of Americans are become adversely affected by different sustenance things every year. In excess of 17 million Europeans have nourishment sensitivities, and clinic affirmations for extreme responses in kids have risen sevenfold over the previous decade, according to the “European Academy of Allergy and Clinical Immunology”.

The allergies are usually caused only by a few chemical substances which can easily be found out. Identification of the food items containing the chemicals can also be used for the purpose. The information can be useful for people to choose, which food item to eat and which one to avoid.

The present system does not clearly help in identifying the items which cause a particular type of allergy. The enhancement that can be made in our project is to group all the food items causing a particular allergy into a group which can in turn help people to avoid items which they are allergic to. It is a vital information as lack of any such system can hamper the health and in turn the growth of a country. The information can be provided to a person through various modes like a mobile application or through a website.

6. References-

[1] Kenneth H. Lai, Suzanne V. Blackley, Li Zhou, “Using Mutual Information Clustering to Discover Food Allergen Cross Reactivity”, 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

[2] Kruti Nargund, Dr. Natarajan S. “Public Health Allergy Surveillance Using Micro-blogs”, 2016 International Conference on Advances in computing, communications and informatics (ICACCI)

[3] A. Velmurugan, T. Ravi “Allergy Information Ontology for Enlightening People”, 2016 International Conference on Computing Technologies and Intelligent Data Engineering(ICCTIDE’16)

[4] Ha X. Dang, Christopher B. Lawrence “Allerdictor : fast allergen prediction using text classification techniques”, PMID 24403538, PMCID PMC3982160

[5] Gennaro Tartarisco, ALlessandro Tonacci, Paola Lucia Minciullo, Lucia Billeci, Giovanni Pioggia, Cristoforo Incorvai, Sebastiano Gangemi “ The soft-computing based approach to investigate allergic diseases: a systematic review”, DOI 10.1186/S12948-017-0066-3

[6] J. Jabez Christopher, H. Khanna Nehemia, A Kannan “A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin test”, Computers in Biology and medicine 65 (2015) 76-84

[7] Raquel Gutiérrez Rivas, Juan Jesús García Domínguez, William P. Marnane, Niall Twomey, Andrey Temko “Real time allergy detection”, 2013 IEEE 8th International symposium on Intelligence signal processing
[8] Kathy Lee, Ankit Agrawal, Alok Chaudhary “Mining Social Media Streams to Improve Public Health Allergy Surveillance”, 2015 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM)

[9] https://www.sciencedirect.com/science/article/pii/S1323893016301137

[10] https://ieeexplore.ieee.org/document/7579262

[11] https://www.researchgate.net/publication/310741790_Disaggregating_Asthma_Big_Investigation_vs_Big_Data

[12] https://en.wikipedia.org/wiki/K-medoids

[13] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[14] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[15] https://en.wikipedia.org/wiki/Support_vector_machine