Project Report for CSE 243
Classification of Autoimmune Diseases

Suzanne B. da Câmara
sdacamar@ucsc.edu
UC, Santa Cruz

Jamie Deng
jdeng31@ucsc.edu
UC, Santa Cruz

Golam Md Muktadir
muktadir@ucsc.edu
UC, Santa Cruz

ABSTRACT

Autoimmune Arthritis diseases are difficult to diagnose in part due to discrepancies between the symptoms outlined for each disease in medical journals and the symptoms experienced by disease sufferers. This project focuses on identifying six Autoimmune Arthritis diseases by their common early symptoms as reported by patients. This would aid the diagnosis process and possibly reduce the time to diagnosis. In order to identify these early symptoms, we analyzed privately held data compiled by the International Foundation for Autoimmune and Autoinflammatory Arthritis (IFAA). It contains early symptoms for the 6 diseases. We employ data cleaning, preprocessing, transformation, and feature engineering on the dataset. Then we use different machine learning algorithms to train models to predict the diseases based on their reported symptoms. Random forest performs the best with an 88% accuracy. We also identify important features for the classification task.

Keywords: Disease Classification, Data Mining for Medical Studies, Early Prediction of Disease with Machine Learning, Autoimmune Disease Prediction, Machine Learning in Health, Machine Learning in Medical Science.

1 INTRODUCTION

1.1 Motivation

We would like to increase the accuracy and shorten the time to diagnosis of sufferers with the following Autoimmune diseases: Ankylosing Spondylitis (AS), Psoriatic Arthritis (PsA), Systemic Lupus Erythematosus (SLE), Sjögren's Syndrome (SS), Adult-Onset Still's Disease (AOSD), and Rheumatoid Arthritis (RA). We hope to accomplish this by identifying the common early symptoms for each disease as reported by the patient.

Autoimmune Arthritis diseases are difficult to diagnose. This difficulty is exacerbated by discrepancies between the symptoms given for each disease in medical journals and the symptoms that patients report experiencing prior to diagnosis. These discrepancies prolong the time to disease diagnosis. It is common for sufferers of the six diseases in the study to go years without an accurate diagnoses. Since these diseases are autoimmune diseases, during this time, the body is literally attaching itself. The more time that lapses between disease onset and disease diagnosis, the more damage occurs. Damage which is irreversible and which reduces the patient's mobility, longevity, and quality of life. Any work in this area that either confirms early symptoms commonly associated with one of these diseases or identifies new ones, aides in the diagnosis process and possibly shortens time to diagnosis resulting in better long term prognosis.

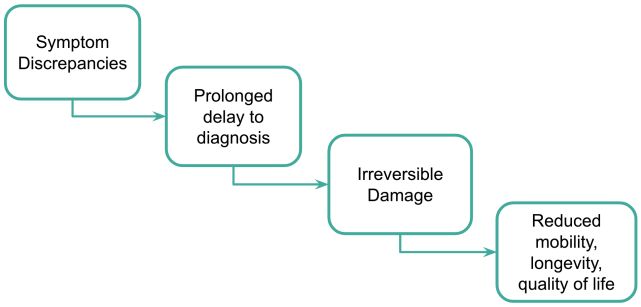

Figure 1: Motivation

1.2 Background

The data in this study was analyzed by Statwing, a data analysis software package associated with Fluid Survey (now Survey Monkey). We expect to be able to confirm

or refute their findings and we hope to be able to gain new insights that can guide future research directions.

2 DATASET

2.1 Overview

We were given access to privately held data that was gathered from a study by the International Foundation for Autoimmune and Autoinflammatory Arthritis. We already have the data in our possession. The data set consists of retrospective patient reported early symptoms for the following 6 diseases: Ankylosing Spondylitis (AS), Psoriatic Arthritis (PsA), Systemic Lupus Erythematosus (SLE), Sjögren's Syndrome (SS), Adult-Onset Still's Disease (AOSD), and Rheumatoid Arthritis (RA). It is comprised of over 900 data objects and 449 attributes. The attributes are a mix of nominal, binary, ordinal, and numeric data.

An initial look at the data revealed that we will need to spend some time cleaning and organizing the data. We will also have to spend considerable time learning the various attributes so that we can collapse the columns into meaningful summaries. For instance, seven columns were devoted to whether or not various family members or extended family members were ever diagnosed with the same disease for each of the 6 diseases resulting in 42 sparsely populated columns. One possible way to collapse the data is to have 2 columns per disease: one for nuclear family and one for extended family. [1]

2.2 Challenges with Dataset

The dataset is built against a set of questions organized as a survey. The accompanying questionnaire has 59 pages. While the meaning of the data may seem obvious to appropriate medical practitioners, the data was far from ready to be analyzed systematically. Here we list some of the major challenges which we tackled with a great effort:

2.2.1 Not all the questions are about earlier symptoms. Some of the questions are related to posterior symptoms or diseases. So, they surface after a person already has developed an autoimmune disease. And as the data-set holds data for those questions, too, it led us to filter those posterior attributes from the dataset. So, first challenge was to confirm which were in fact early symptoms.

2.2.2 Mapping from questions to data object attributes were not obvious. For this reason, we couldn't readily tell which column refers to which question. This made our first challenge even more difficult. In the early experiments, we inadvertently included some posterior attributes for prediction models and ended up very high accuracy rates but unfortunately wrong. To make the mapping clear, we renamed all the columns in such a way that we could find the associated question fast. But that was not enough as many of the questions didn't have question numbers. So, for some questions, we created unique prefixes which can group related questions. We also made the prefixes to hold some semantics. For example, we use the prefix **Post_AD_**, for all the attributes which relates to diseases happening after the first autoimmune disease. Smart prefixing also helped us to filter out attributes automatically. So, for first disease prediction, we matched the column names with Post_AD_ and AAD_AD_ prefixes and dropped them. Without renaming columns it would be a tiresome work for us to manipulate columns for transformations and clean-up.

2.2.3 Inconsistent attribute values. When it comes to attribute values for a real-life survey, all kinds of surprises are common. Here are some that we experienced and how we handled it:

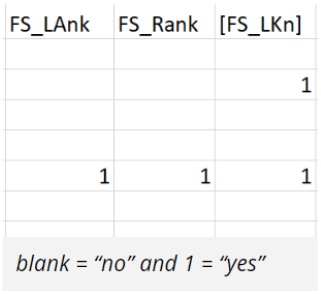

Figure 2: Binary attributes

- For some attributes, an empty string denotes "no" or 0, if they are binary valued (Figure 2)
- Different attributes have different implicit definitions for **Missing Values**. Columns in Figure 3, both a empty string and "I do not remember" denote missing values.
- For some attributes, there are values that can only be fixed manually. For example, the Onset_Age column is numeric, but it can have non-numeric

values as shown in Figure 4. As the survey tool did not validate the form data, end users are able to enter anything they want.

[LHand_ABCJoint_0_12]	[RHand_ABCJoint_0_12]
The larger joints only were affected	Both the small and larger joints were affected
Both the small and larger joints were affected	Both the small and larger joints were affected
I do not remember	I do not remember

blank = "missing",
I do not remember = "missing"
Others = categorical values

Figure 3: Categorical attributes

50 ?
30
24
22
Unsure

Onset_Age column

Figure 4: Numeric column

2.2.4 Attributes having values of both Pre and Post symptoms. Some attributes can have values from both early symptoms and from posterior symptoms. For example, in the question 5, the first two values are early symptoms, but the third value is posterior. We created two new attributes from the data of this attribute, one having values related to early symptoms only and the other having all with the second value transformed as a "no".

12. Have you ever been diagnosed with fibromyalgia?

- ○ Yes, I was diagnosed with fibromyalgia prior to my Autoimmune Arthritis diagnosis and I do, in fact, have fibromyalgia in addition to my Autoimmune Arthritis disease.
- ○ Yes, I was diagnosed with fibromyalgia prior to my Autoimmune Arthritis diagnosis, however, it was later determined that I do not have fibromyalgia.
- ○ Yes, I was diagnosed with fibromyalgia after my Autoimmune Arthritis diagnosis. Please specify month/year of diagnosis: _____
- ○ No, I have never been diagnosed with fibromyalgia.

Figure 5: Values for attribute fibromyalgia

2.2.5 Invalid utf-8 characters. Last but not the least and probably the most difficult problem to solve was to fix invalid characters. We manually did that in Microsoft excel where they showed up as weird whitespaces.

3 METHODS AND PIPELINE

3.1 Data mapping & Attribute Renaming

The pivotal element of a smart mapping between the questions and dataset attribute is the **Prefixing mechanism** we employed. The prefixed we created served several purposes:

- To denote which section, group, or question the attribute belongs to. For example, AAD_AD prefix denotes the group of the chosen 6 autoimmune diseases which can develop after the first autoimmune disease.
- Some questions has multiple choice options. Each of the option becomes a attribute in the dataset. So, they were prefixed with the same string. This helps to filter out the whole question from dataset.
- Prefixes also help us to segment data. Though, we did not conduct any experiment with segmentation in this project, it will be helpful for future work.

In addition to help us to build effective mapping, renaming also helped us to create transformers easily. For example:

Example 1. "[M_Thr_Nk_Ns_24] 34. Did you have any of the following issues with your mouth/ throat/ neck/ nose region during the first 24 months after initial onset? [Painful swollen and tender lymph nodes in areas of the body not including the face and/or neck?]"

is a actual attribute name in the raw dataset. If we want to refer to the column in our automated cleaning and transformation scripts, that would be cumbersome to work with. But our new name for the column, **Ex5_34_body**, was easy to refer. Here Ex5 denotes section of "Exhibit 5", 34 denotes "Question no 34", and body denotes the option with body excluding face and neck.

3.2 Data Cleaning & Transformations

We call automated transformation algorithms by **transformers**. Before transforming values of an attribute, we need to make sure they are clean and consistent enough for the transformer to work. Each transformer can apply a set of ordered basic transformers.

For **continuous values** like age, we automated detection of non-numeric values and manually fixed them in Microsoft Excel. For **categorical values** we followed the following steps:

(1) Figure out definitions of missing values for the attribute
(2) Replace missing values with string "NA2". "NA2" was chosen instead of "NA" as some analytic and machine learning libraries automatically processed "NA" as missing values. We wanted full control on our processes.
(3) Figure out unique values. This itself is two step process. First step is automatically extract unique string values. Then manually detect "duplicates" and map duplicates to a single value.
(4) Create a transformer to transform all the values for the attribute.

For **mixed attributes** having data from both pre and post symptoms, we needed to extract new attributes before applying the transformations.

3.3 Prediction

We followed a 3-step process for prediction task of first autoimmune disease:

(1) **Attribute selection for objective:** First we selected the appropriate columns for our objective. We removed some attributes automatically based on **Prefix Search**. Then we manually dropped attributes which had no relation with our objective.
(2) We build some models with the selected attributes. Dropping missing values was based on type of model we were using. As our dataset is small, for some models, we dropped columns instead of rows.
(3) Compare results.

3.4 Final Pipeline

This is a final pipeline of our whole process. It was developed through process iterations in our experiements.

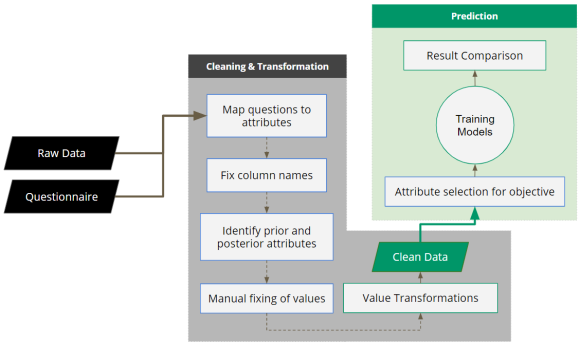

Figure 6: Final process pipeline. The blue-bordered methods requires manual interventions and optionally some automated processes.

4 EXPERIMENT SETTING

4.1 Configurations:

- Split the dataset into training set and test set, training set contains 80% of data instances, test set contains the rest 20%.
- We use Random Forest, neural network and SVM, RF performs the best, SVM worst.
- Random Forest is also used to identify which symptoms are important for each disease.

4.2 Technologies used

- **Python** was used to write all the transformers and final pipelines.
- **Scikit-Learn** was used to develop all the final models
- **Microsoft Excel** was used for question attribute mapping, manual value fixes
- **Rapidminer** was used to find duplicate column names, correlations, initial model development, and initial attribute significance analysis.

4.3 Experimental pipeline

Our final process pipeline 6 emerged from experimental pipeline in 7. We revised the cleaning and transformation processes several times which resulted in a

sequence of revisions in and execution of other processes.

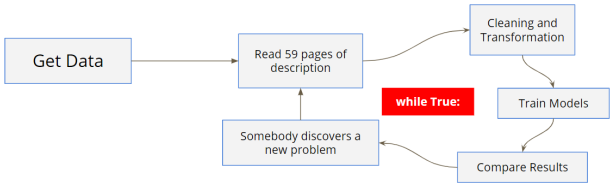

Figure 7: Experimental pipeline

5 RESULTS AND DISCUSSION

For training the model, we use a number of different machine learning methods including Random Forest, Neural Network, and Support Vector Machine (SVM). We draw those methods from `Python Scikit-Learn` library [3]. Random Forest doesn't require many parameters, the most important one is the number of trees n. [2] suggest the number between 64-128. We choose $n = 100$ because the model would converge and more trees wouldn't improve the accuracy much. The parameters for Neural Network and SVM are selected by manually tuning the algorithms to get better accuracy. For neural network, we select 3 hidden layers, each with 60 nodes. Since the dataset is not very large, more layers or nodes would cause overfitting on the training dataset. For SVM, we select $\gamma = 0.01$ and $C = 10$.

5.1 Classification accuracy

Random Forest shows the highest accuracy of 88% in classifying different diseases. Neural Network performs the second with accuracy of 84%. SVM only has 74% accuracy. The following figures show the performance metrics for each algorithms. The results also show performance scores for each class (disease). Overall, Random Forest performs the best. Random Forest usually has the best scores predicting different classes except for psoriatic. Neural Network outperform Random Forest on predciting psoriatic especially with high recall score.

	precision	recall	f1-score
ankylosing	0.82	0.94	0.88
psoriatic	0.73	0.79	0.76
rheumatoid	0.86	0.82	0.84
sjogren	0.90	0.75	0.82
still	1.00	0.88	0.94
systemic	0.94	0.98	0.96
accuracy			0.88

Figure 8: Random Forest, performance metrics

	precision	recall	f1-score
ankylosing	0.86	0.89	0.87
psoriatic	0.72	0.93	0.81
rheumatoid	0.84	0.79	0.82
sjogren	0.77	0.71	0.74
still	1.00	0.82	0.90
systemic	0.87	0.90	0.88
accuracy			0.84

Figure 9: Neural Network, performance metrics

	precision	recall	f1-score
ankylosing	0.82	0.80	0.81
psoriatic	0.83	0.71	0.77
rheumatoid	0.61	0.79	0.69
sjogren	0.61	0.46	0.52
still	1.00	0.53	0.69
systemic	0.80	0.88	0.84
accuracy			0.74

Figure 10: SVM, performance metrics

SVM doesn't give good prediction on some of the diseases, especially sjogren. Also, its recall scores on sjogren and still are very low which affect its overall performance.

5.2 Feature importance

We also use Random Forest to identity which symptoms or features are important for predicting and for each disease. Random Forest has a build-in feature importance attribute which is computed during training by recording impurity scores calculated when splitting

the trees. The top ranking most important features include: antinuclear antibodies, HLA-B27 gene, rheumatoid other symptoms, sjogren other symptoms, onset age, psoriatic other symptoms, country, and so on.

The top five ranking features for each disease are:

- still:
(1) antinuclear antibodies
(2) HLA-B27 gene
(3) degree of fever during first 24 months
(4) rheumatoid other symptoms
(5) skin rashes

- systemic:
(1) antinuclear antibodies
(2) HLA-B27 gene
(3) rheumatoid other symptoms
(4) country
(5) diagnosed with Psoriasis

- ankylosing:
(1) antinuclear antibodies
(2) HLA-B27 gene
(3) rheumatoid other symptoms
(4) sjogren other symptoms
(5) pain/tenderness

- rheumatoid:
(1) antinuclear antibodies
(2) HLA-B27 gene
(3) onset age
(4) sjogren other symptoms
(5) swelling in pain/tenderness locations

- psoriatic:
(1) antinuclear, diagnosed with Psoriasis
(2) antibodies
(3) HLA-B27 gene
(4) rheumatoid other symptoms
(5) psoriasis begin before your other symptoms

- sjogren:
(1) antinuclear antibodies
(2) HLA-B27 gene
(3) rheumatoid other symptoms
(4) onset age
(5) country

By computing the feature importance, we could identify for each disease which features/symptoms play more significant parts in predicting the diseases. Thus enhance our knowledge in the domain. The deciding features for all diseases are very similar, we still have much to learn.

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

Autoimmune Arthritis are difficult to diagnose because of discrepancies between the symptoms outlined for each disease in medical journals and the symptoms experienced by disease sufferers. This project focuses on identifying 6 of these diseases by their common early reported symptoms. We analyzed privately held dataset, which contains the symptoms, compiled by the IFAA. We employ data cleaning, preprocessing, transformation, and feature engineering on the dataset. Then we use different machine learning methods to train models in order to predict the diseases. We create a model which can differentiate between these 6 diseases with 88% accuracy by using Random Forest. We also identify important features for the classification task.

6.2 Future work

Possible future work includes:

(1) Create survey that lends itself to analysis and that also includes questions that will help us answer questions raised during analysis. Identifying patterns of symptoms over time for different diseases/segments.
(2) Segmentation Analysis
(3) More experiments with models
(4) Predicting a sequence of Autoimmune diseases
(5) Identifying relevant symptoms

REFERENCES

[1] IFAA. Early symptoms of AiArthritis study.

[2] Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (2012), Springer, pp. 154–168.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, Oct (2011), 2825–2830.