Diagnosis Of Erythemato-Squamous Diseases

Achint[1]

[1]Affiliation not available

January 13, 2018

Abstract

Dermatological diseases need a better classification system than just the clinical trials. This is further supported by the fact that most of the symptoms are at a cellular level. Moreover, the diseases are more often than not correlated and may show the symptoms of another disease at a beginning stage. This need of better classification has led to a need for research in this field.

Introduction

In today's world, correct diagnosis of disease has become an essential part in medical science. Incorrect diagnosis of diseases lead to endangerment of lives on a daily basis. So, computer based diagnostic machine systems can play an important role in accurate diagnosis of diseases. Also, there is a huge availability of data of patients - from the initial reports to the treatments to the prescriptions and the follow-ups.

However, there is a very improper organization of huge amount of data which is affecting the quality of decision making. This increase in the volume of bulk of data requires some way in which it can be organized, extracted and processed efficiently.

Health care industry today provide a lot of applications like treatment effectiveness, Health-Care management, Customer Relationship management and Pharmaceutical management among other things. One such application is the discovery of patterns and relationships among clinical and diagnostic data using machine learning techniques.

In the past decade, machine learning has paved the way for a lot of features like self-driving cars, practical speech recognition, effective web searches, and a vastly improved understanding of the human genome. They are data driven approaches, mainly designed to discover statistical patterns in high dimensional, multivariate data sets, frequently found in electronic health records. Pattern identification in machine learning makes it a powerful tool for predictions and decision making process for diagnosis and treatment planning.

The classification learning algorithm in most of diagnostic cases are composed of two main components : training and classification. The training phase involves a formation of model based on some previously trained examples of that domain. The classification phase, using this model, tries to predict the class that a new instance belongs to. The main requirement of such a system is prediction accuracy. The time taken by such a system to predict classes should also have a short training and prediction time. Such a system should also be robust to noisy training instances. Both training and test instances may have some missing values. That should also be dealt with by that system. The features that are used to encode the instances may also

have different levels of relevancy to the domain. There are many more factors that needs to be dealt with by that system.

Moreover, the use of classification algorithm in case of diagnosis of diseases is two fold. First, the actual doctors can check and verify the learned classification knowledge before it is deployed into the real world. Moreover, it may shed some light on some previously unknown pattern or fact, leading to newer discoveries in the field.

Diagnosis of Dermatological Diseases

The diagnosis of erythemato-squamous diseases is a tricky problem in dermatology. Most of the symptoms are correlated to each other, with very few differences. The diseases to be classified are *psoriasis, seboric dermatitis, lichen planus, pityriasis rosea, chronic dermatitis* and *pityriasis rubra pilaris.*

To the naked eye, these diseases look very much alike to *erythema* and *scaling.* But when inspected more carefully, some patients have the typical features of a particular disease at typical localizations (skin level) while some features of that disease might be at the histopathological level (cellular level).

The patients were first evaluated clinically with 12 features - the degree of *erythema* and *scaling*, whether the *borders of lesions* are definite or not, presence of *itching* and *koebner phenomenon*, the formation of *papules*, whether the *oral mucosa, elbow, knees* and *scalp* are involved or not, and whether there has been instances of these diseases among the patient's predecessors.

Initial univariate and bivariate analysis of the data showed some patterns. First off, the level of *erythema* and *scaling* of *chronic dermatitis* is much less than in the case of *psoriasis.* Moreover, the *koebner phenomenon* is present only in *psoriasis, lichen planus,* and *pityriasis rosea. Itching* and *polygonal papules* are mainly found in *lichen planus.* However, *follicular papules* are mainly the cause for *pityriasis rubra pilaris.* While the involvement of *elbow, knees* and *scalp* affect *psoriasis* more, *oral mucosa* is a predilection site for *lichen planus.* Family history is generally present for *psoriasis* while *pityriasis rubra pilaris* usually start during childhood.

Usually, patients are generally diagnosed using these clinical features. However, a biopsy is necessary for correct and definite diagnosis.

So, skin samples were taken for evaluation of 22 histopathological features. The main difficulty in analyzing histopathological features is that a disease may show the features of one type of disease at one stage and features of another type of disease at a later stage. Some may not even show the features of a particular disease they are having. Some features like *acanthosis* and *parakeratosis* can be found in all of the diseases at different levels. While other features may be particular to one particular type of disease, like *melanin inconsistence* for *lichen planus.*

Based on the statistical analysis done by looking at the univariate and bivariate plots, it can be said that few clinical symptoms like *koebner phenomenon* and *follicular papules* were mainly found in a particular disease rather than in all the diseases. However, many histopathological symptoms like *melanin inconsistence* are particular to only one type of disease in their respective cases. This exclusivity is much higher in the case of histopathological features than in the case of clinical symptoms. This can be summarized by an

observation that histopathological features can be used for a better classification of these dermatological diseases than the clinical trials held for observation.

Dataset Description

The dataset has records of 366 patients. 12 clinical features and 22 histopathological features are used as features in the training of models. The family history attribute is a categorical attribute that has a value of 1 if any one of these diseases was present in the predecessor, else it has a value of 0. The age attribute is a numerical attribute which simply represents the age of each patient. Rest of the features are categorical and range from values of 0 to 3.

Here: 0 - Symptom not present in the patient

1 - Symptom present in small amounts

2 - Symptom present in moderate amounts

3- Symptom present in large amounts

Modeling

First, the data was prepared by changing the attributes to their particular data types. Then, the missing values were imputed by using missForest package. It uses a random forest trained on the observed values of a data matrix to predict the missing values. It can be used to impute continuous and/or categorical data including complex interactions and non-linear relations. It yields an out-of-bag (OOB) imputation error estimate without the need of a test set or elaborate cross-validation. It can be run in parallel to save computation time.

Multinomial regression is first applied to the dataset as an initial benchmark. The basic intuition behind it is using log of odds ratio by doing one-vs-all method. Using multinomial regression method resulted in the following results :

	Accuracy	Precision	Recall	F1_Score
Class 1	94.23	100	88.46	93.87
Class 2	86.74	85.71	75	80
Class 3	100	100	100	100
Class 4	98.41	84.61	100	91.66
Class 5	99.20	91.66	100	95.65
Class 6	99.27	83.33	100	90.90

A basic C5.0 decision technique was also used for classification. It performed slightly better than multinomial regression.

	Accuracy	Precision	Recall	F1_Score
Class 1	98.07	100	96.15	98.03
Class 2	93.75	100	87.50	93.33
Class 3	99.18	92.85	100	96.29
Class 4	100	100	100	100
Class 5	100	100	100	100
Class 6	99.27	83.33	100	90.90

CART model was also used for classification as it generally favors the use of categorical attributes. Hyper parameter tuning had little effect on the results of the initial model.

	Accuracy	Precision	Recall	F1_Score
Class 1	96.15	100	92.30	96
Class 2	91.47	70	87.50	77.78
Class 3	100	100	100	100
Class 4	95.45	100	90.90	95.23
Class 5	100	100	100	100
Class 6	99.27	83.33	100	90.90

A Bagging CART model was also generated. It creates a set of decision trees and bags the average prediction of each instance. Similar to CART, it also favors the categorical variables.

	Accuracy	Precision	Recall	F1_Score
Class 1	100	100	100	100
Class 2	98.48	80	100	88.89
Class 3	100	100	100	100
Class 4	90.90	100	81.81	90
Class 5	100	100	100	100
Class 6	100	100	100	100

Gradient Boosting was also implemented. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differential loss function.

	Accuracy	Precision	Recall	F1_Score
Class 1	100	100	100	100
Class 2	98.36	86.66	100	92.85
Class 3	100	100	100	100
Class 4	85.71	100	71.42	83.33
Class 5	100	100	100	100
Class 6	100	100	100	100

Random Forest was also implemented. Random forests are an ensemble learning method for classification other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

	Accuracy	Precision	Recall	F1_Score
Class 1	100	100	100	100
Class 2	99.18	92.85	100	96.29
Class 3	100	100	100	100
Class 4	92.85	100	85.71	92.30
Class 5	100	100	100	100
Class 6	100	100	100	100

Although machine learning techniques like decision tree models (rpart and C5.0) and gradient boosting performed well, random forest gave the best result. Multi-layered perceptrons could also be implemented but it was not a good approach as the number of instances were quite low. Dimensionality reduction and feature engineering were also not done as each and every symptom is as important as the next one, however rare it may be.

Conclusion

As it turns out in the statistical analysis, histopathological features can be used for a much better classification of diseases than just the clinical trials. Also, most of the models can be implemented as part of a classification engine but since better classification of each type of disease is a priority, the most apt engine would be that using random forest algorithm.