ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Assessment of diabetic retinopathy risk with random forests

Silvia Sanromà[1], Antonio Moreno[1], Aida Valls[1], Pedro Romero[2], Sofia de la Riva[2] and Ramon Sagarra[2]*

[1]Departament d'Enginyeria Informàtica i Matemàtiques – Universitat Rovira i Virgili
Av.Països Catalans, 26. 43007-Tarragona - Spain
[2]Hospital Universitari Sant Joan – Universitat Rovira i Virgili
Av. Dr. Josep Laporte, 2. 43204-Reus - Spain

Abstract. Diabetic retinopathy is one of the most usual morbidities associated to diabetes. Its appropriate control requires the implementation of expensive screening programs. This paper reports the use of Random Forests to build a classifier which may determine, with sensitivity and specificity levels over 80%, whether a diabetic person is likely to develop retinopathy. The use of this model in a decision support tool may help doctors to determine the best screening periodicity for each person, so that an appropriate care is provided and human, material and economic resources are more efficiently employed.

1 Introduction

Diabetes Mellitus (DM) is one of the more prevalent chronic diseases in the world. According to the World Health Organization, 347 million people worldwide (around 4.6% of the population) suffer from DM, and it has been predicted that it will be the 7th cause of death by 2030. Only in 2012 it was the direct cause of 1.5 million deaths†. It is also a leading cause of complications such as blindness, amputation and kidney failure. *Diabetic retinopathy* (DR) is one of its more widespread morbidities and it has been increasing steadily in the last years. Its main effect, secondary blindness, has a large social and economic impact in healthcare. The early detection of DR, by means of periodic controls, reduces significantly the financial cost of the treatments and decreases the number of patients who develop blindness [1].

Some scientific societies recommend that diabetic patients should be screened for DR every year‡; however, in practice this periodicity is very hard to achieve, due to the large number of diabetic people, the lack of enough human and material resources in medical centres and the economic cost of the screening procedure. Thus, there is a strong interest in developing a tool that can analyze the personal and clinical data of a diabetic person and help the medical practitioner to determine his/her risk of

* This study was funded by the research projects PI12/01535 and PI15/01150 (Instituto de Salud Carlos III) and the URV grant 2014PFR-URV-B2-60.
† http://www.who.int/features/factfiles/diabetes/facts/en/
‡ For example, the American Diabetes Association [2], the American Academy of Ophthalmology and the Royal College of Ophthalmologists [3].

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

developing DR, so that the temporal distance between successive controls may be adjusted depending on it and human and material resources may be used more efficiently.

In the last year researchers especialised in Ophthalmology and Artificial Intelligence at University Rovira i Virgili have been working on the application of *Intelligent Data Analysis* techniques to data from diabetic patients in order to develop a model that may predict whether a certain person is likely to suffer DR. Several classification techniques have been analyzed, including *k-Nearest Neighbours*, *Decision Trees* [4] and *regression functions*. This paper reports the results obtained with a classification model based on *Random Forests* (RF) [5].

The rest of the paper is organised as follows. The next section describes the data that have been analyzed and how the general RF method has been adjusted to the particularities of this problem. Section 3 presents the results of the classification model and compares it with the baseline classification mechanism used until now in the hospital, based on regression. The final section includes the main conclusions and the lines of future work.

2 Material and methods

2.1 Data from diabetic patients

A set of real patient data, including 1743 diabetic people that had not developed DR and 579 that suffered the disease, was provided by the ophthalmologists from Sant Joan Hospital (Reus). In order to test some classification methods this set was randomly divided into a training set T (871 healthy people, 341 people with DR) and a validation set S (872 healthy, 238 with DR). From now on, the class of healthy patients will be called 0 and the one of DR individuals will be called 1. Thus, the aim of the work was to develop a data analysis procedure that, after analyzing the data from set T, could build a classification model that could predict accurately whether the individuals in set S belong to class 0 or to class 1.

Each individual is described by 9 attributes including personal characteristics (e.g. age, gender) and clinical data (e.g. hypertension). The attributes were determined after the analysis of a period of 8 years on a population of 17000 diabetic patients. This study identified the attributes with stronger influence on the risk of having retinopathy [6]. The attributes were continuous or categorical. The continuous ones (e.g. age) were divided into relevant intervals according to [6], so that all attributes were finally treated as categorical. The values of these attributes were taken at the moment of the diagnosis of DR.

2.2 Classification model based on a Random Forest

Given a set of pre-classified objects defined on a set of categorical attributes, the algorithms for inducing decision trees (e.g. ID3 [4]) build a hierarchical structure that allows classifying any other object. In a decision tree each node represents an attribute, and the children of the node are labelled with the attribute values. The leaves of the tree indicate the class to which an object with the values shown in that branch belongs. These algorithms assume that all the objects in the training set that

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

share the same values in all the attributes belong to the same class, as they are indistinguishable. This fact presents a problem in our case, since it may be the case that, given a value associated to each of the 9 attributes, some patients in T belong to class 0 and others to class 1. More concretely, taking into account all the possible values of the 9 attributes there are 4608 combinations. The training set, containing 1212 individuals, only contains 451 of those combinations. Moreover, in 120 of them there are patients from both classes. In order to deal with this issue without losing information we keep the number of patients in each class for each combination. This number is used to assign a class (0, 1 or unknown) to each leaf of the decision tree.

While a decision tree has many advantages, such as comprehensibility and scalability, it still suffers from several drawbacks—instability, for instance. One way to realize the full potential of decision trees is to build a decision forest [7]. In the Random Forest method several decision trees are constructed and the final decision takes into account the predictions of all the trees. Here is how each tree of the random forest is obtained from a training data set T:

1. Pick up randomly N items of the training data. Some studies suggest that N should be around two thirds of the training set [5]. As we want to have a balanced set of items to build each tree, we take 340 patients from each of the two classes, for a total number of 680 items (56% of the training set).
2. At each node:
    a. *m* attributes are randomly selected from all the ones that have not been used yet in that branch. Previous works suggest that this number should be around *log*(number of attributes) [7].
    b. The entropy of each of these attributes is computed to determine the one that classifies better the training examples remaining in that branch and we create successors nodes for each of its values. The process stops (and a leaf of the tree is created) when, considering the combinations covered by the branch, the percentage of individuals of the training set from one class exceeds a given threshold. An "unknown" label is given to a leaf if there are not any more attributes to consider and none of the two classes exceeds the threshold.

3 Experimental setting

In this section it is explained how the optimal values for the parameters of the Random Forest method were determined. After that, the results of the RF classification are compared with those given by other well-known methods. In all the tests described in this section the following evaluation measures were considered:

- Sensitivity: TP/ (TP+FN)
- Specificity: TN/(TN+FP)
- Accuracy: (TP + TN)/ (TP+TN+FP+FN)

TP are True Positives (class 1, prediction 1), FP are False Positives (class 0, prediction 1), TN are True Negatives (class 0, prediction 0) and FN are False Negatives (class 1, prediction 0). Our aim was to obtain a classification method with sensitivity over 80%, as required for example by the British Diabetic Association.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

3.1 **Random Forest parameters**

The standard RF technique has two basic parameters: the number of trees of the forest and the number of attributes considered in each node. Moreover, in our case it is also necessary to determine the value of the threshold that controls the creation of the leaves of the tree.

Let us start the analysis with this threshold. Tests were made with values between 60% and 95%, taking 200 trees in the forest and 2, 3 and 4 attributes in each node. All results show that 68% is the optimal value. Figure 1 shows the sensitivity, specificity and accuracy of the resulting RFs for 2 attributes. We can see that with a value of 68, the three evaluation measures are closer to 80%.With a higher value it is possible to increase specificity keeping a good accuracy, but there is a very strong decrease in sensitivity.

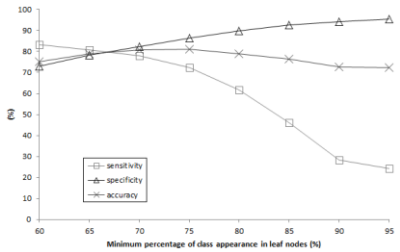

Fig. 1: Analysis of the leaf-creation threshold

On the second place we studied the influence of the number of attributes considered in each node of the tree. In the tests we tried the values from 1 to 4, with the threshold 68% and 200 trees in the RF. The results (Figure 2, left) show that 3 is the unique value for which the three evaluation measures exceed 80%.

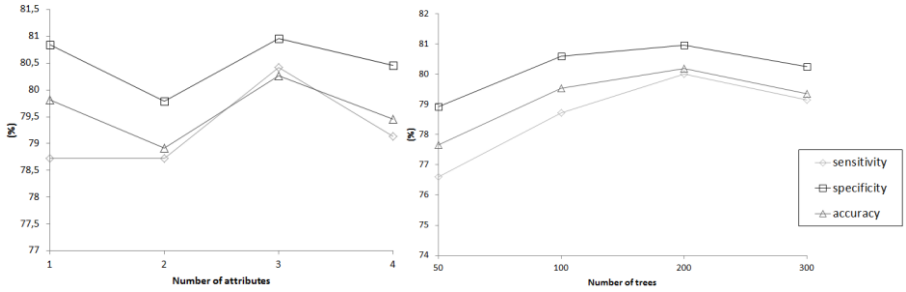

Fig. 2: Analysis of the number of attributes and the number of trees.

Finally, the influence of the number of trees in the RF was analyzed, taking the 68% threshold and 3 attributes in each node. The values considered in the study were 50, 100, 200 and 300. As seen in Figure 2 (right), the best performance of the three evaluation measures was reached when 200 trees were considered. In summary, the final RF setting considered 200 trees, 3 randomly selected attributes in each node and a minimum leaf-creation percentage of 68%.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

3.2 Classification results

For each element of the validation set S the 200 trees are used to obtain 200 predictions of the class of the element, which may be 0, 1 or unknown (in some cases a tree may fail to classify an object because it lacks the branch with the attribute value in a given node or because there are no attributes left to explore and none of the classes has reached the required threshold). The element is assigned to the class with a higher number of predictions. If there is a tie in the number of predictions, the preference is in the following order: unknown, 0, 1.

		Predicted class			
		0	1	unk.	
Real class	0	702	165	5	Specificity: 80.96%
	1	47	188	3	Sensitivity: 80.00%

Table 1: Classification using Random Forest

Table 1 shows the classification results. It may be seen that the system is able to make a prediction in almost all of the cases (it only fails to make a prediction in 8 out of 1110 patients, 0.72%). The values of specificity and sensitivity reach 80%, whereas the global accuracy of the predictions is 80.76% (890/1102).

3.3 Comparison with other methods

We have compared the results of the system with three other well-known classification methodologies, given below. In the two first methods below, the dataset was previously balanced (replicating patients with RD) so that they can be fairly compared with Random Forest. However, in the last method, we used a non-majoritary prediction technique that internally manages the imbalance between the number of cases in each class.

- *Logistic regression*: this is the classification method used by the ophthalmologists of the hospital before the start of the research reported in this paper, so it can be taken as the reference baseline. A statistical package was used to calculate the regression function with a Logit model, 95% of confidence interval, 100 iterations, 0.000001 of convergence and using the Newton-Raphson algorithm for the maximization of the likelihood function.
- *Decision tree*: we built a decision tree from all the data of the training set T using the classical ID3 algorithm [4]. A leaf is introduced in the tree when the percentage of individuals belonging to a class (from all the individuals considered in that branch) exceeds 89%. This number was empirically found to be the one that leads to better classification results.
- *k-Nearest Neighbours*: for each patient of the validation set S, we look for the 5 patients in the training set T that are more similar. The similarity measure between two patients is the addition, for all the attributes, of the difference between the attribute values of the two patients divided by the number of possible values of that attribute. The best results of this method appear when the system predicts class 1 if at least one of the five neighbours belongs to class 1 (i.e. class 0 is predicted only if the five neighbours belong to class 0).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

	Regression	ID3	k-NN	RF
Sensitivity	51.42%	60.08%	25.21%	80.0%
Specificity	94.49%	66.78%	77.52%	80.96%

Table 2: Comparison of the sensitivity and specificity of the classification methods

In Table 2 it may be seen that the regression function provides a high specificity (almost no False Positives), but the sensitivity hardly exceeds 50%. The k-NN method has specificity close to 80%, but sensitivity is too low (25%). ID3 achieves a similar sensitivity and specificity, but they are also too low (below 70%). In general, the main problem of the three methods is that they give a very high number of predictions for the class 0 even if the data is balanced. Thus, the number of false negatives is too high to be acceptable because many patients with risk of developing DR are not detected.

4 Conclusion and future work

Doctors need to be able to predict accurately which patients have a high risk of developing diabetic retinopathy, so that the limited human, temporal and material resources available in the screening programs are efficiently used. Thus, classification methods with a high sensitivity are required. Some standard techniques like regression functions, single decision trees or k-nearest neighbors do not have a good performance in this problem, the main reason being the inherent uncertainty in clinical data (patients with the same characteristics may appear in both classes). As seen in this paper, Random Forests provide a good classification model, obtaining sensitivity and specificity values over 80%. In our current work we are studying the relationship between the certainty in the prediction (the percentage of the majoritary class) and the cases with a classification error (False Positive or False Negative). A study of the rules that give the best performance and the key attributes is also planned. On the medium term, our aim is to introduce this classification model in a decision support tool in Primary Care to help doctors decide whether to send a patient to an ophthalmologist for a more detailed examination. Moreover, we want to extend the model to predict also the DR severity (which is classified in 4 levels).

References

[1] J.S.Edwards. Diabetic retinopathy screening: a systematic review of the economic evidence. *Diabetic Medicine*, 27 (3): 249-256, 2010.

[2] American Diabetes Association. Standards of medical care in diabetes. Microvascular complications and foot care. *Diabetes Care* 38:S58-S66, 2015.

[3] The Royal College of Ophthalmologists. Diabetic Retinopathy Guidelines, 2012. (rcophth.ac.uk)

[4] J.R.Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986.

[5] L.Breiman. Random forests, *Machine Learning* 45: 5-32, 2001.

[6] Romero-Aroca P, de la Riva-Fernandez S, Valls-Mateu A, Segarra-Alamo, R., Moreno-Ribas, A., Soler, N., Changes observed in diabetic retinopathy: eight-year follow-up of a Spanish population, *Br J Ophthalmol* Published Online: 14th January 2016. doi:10.1136/bjophthalmol-2015-307689

[7] L.Rokach. Decision forest: twenty years of research, *Information Fusion*, 27: 111-125, 2016.