CSP 571: Project Proposal and Outline

Project Members:

1.Rahul Bharadwaj Machiraju (A20502085) - Team Leader
2. Sai Pavan Kunda(A20496516)
3. Rahul Maddula(A20488730)

1 Project Proposal

1.1 Formal description and stated research goal

The thyroid is an endocrine gland located in the anterior region of the neck: its main task is to produce thyroid hormones, which are functional to our entire body. Its possible dysfunction can lead to the production of an insufficient or excessive amount of thyroid hormone. Predictions about the treatment can be important for supporting endocrinologists' activities and improve the quality of the patients' life.

Specific questions :

1) What are the various features in the complete dataset which contribute the most to predict the type of disease?
2) Out of all the target classes in the dataset, which are the most commonly occurring one's?
3) Which features contribute the most to predict the disease from the selected target classes?
4) Which among the k-NN and Neural Networks predicts the data best?

Methodology :

1.)Firstly, we need to identify the major target classes and prepare the data accordingly.
2.)Clean the data based on the target classes and arrange the data accordingly.
3.)Then perform exploratory data analysis on the data to identify the key features.
4.)After identifying the key features, split the data into train and test. Use the training data for training the designed models.
5.)After training, test the trained model with the testing data and see how well it performs.
6.)Finally, find the Accuracy ,precision and f-1 score of the models and conclude which among the two used classifiers perform well on the data and state reasons why.

Set of metrics which will measure analysis results.

- Find the Accuracy, precision and f-1 score of the models and conclude which among the two used classifiers perform well on the data and state reasons why.

- Root mean square error is a commonly used statistic for determining the disparities between values predicted by a model and the values actually observed.

2 Project Outline

2.1 Literature review and related work

Reviewing the references in the reference section can help you understand the problem's context.

2.2 Data Source and References

2.2.1 Data
Data Source: Kaggle
Dataset Type: CSV file

Overview:

The dataset is from a Kaggle competition which contains 9172 rows and
31 columns.

2.2 Feature Description

- **The details of the dataset, from Kaggle, are explained below:**

Column Name	Datatype	Description
age	int	Age of the patient
sex	str	sex patient identifies
on_thyroxine	bool	whether patient is on

		thyroxine
query_on_thyroxine	bool	whether patient is on thyroxine
on_antithyroid_meds	bool	whether the patient is on antithyroid meds
sick	bool	whether patient is sick
pregnant	bool	whether patient is pregnant
thyroid_surgery	bool	whether patient has undergone thyroid surgery
I131_treatment	bool	whether patient is undergoing I131 treatment
query_hypothyroid	bool	whether the patient believes they have hypothyroid
query_hyperthyroid	bool	whether the patient believes they have hyperthyroid
lithium	bool	whether patient * lithium
goitre	bool	whether patient has goitre
tumor	bool	whether patient has tumor
hypopituitary	float	whether patient * hyperpituitary gland
psych	bool	whether patient * psych
TSH_measured	bool	whether TSH was measured in the blood
TSH	float	TSH level in blood from lab work
T3_measured	bool	whether T3 was measured in the blood
T3	float	T3 level in blood from lab work
TT4_measured	bool	whether TT4 was measured in the blood

TT4	float	TT4 level in blood from lab work
T4U_measured	bool	whether T4U was measured in the blood
T4U	float	T4U level in blood from lab work
FTI_measured	bool	whether FTI was measured in the blood
FTI	float	FTI level in blood from lab work
TBG_measured	bool	whether TBG was measured in the blood
TBG	float	TBG level in blood from lab work
referral_source	str	
target	str	hyperthyroidism medical diagnosis
patient_id	str	unique id of the patient

2.3 Data processing

The data taken from Kaggle website has both continuous and discrete data which undergoes pre-processing. The missing value and not a number constraint are checked. If the missing value or Not a Number values are present and it is replaced by the mean value of the column. And also The class counts clearly show that the dataset is highly imbalanced. For instance, most of the samples in the dataset do not belong to any particular class. Therefore, the data preprocessing is performed to obtain the standard dataset for our performance evaluation.

The table describes the class-wise response

Condition	Diagnosis class	Count
hyperthyroid	hyperthyroid (A)	147
	T3 toxic (B)	21
	toxic goiter (C)	6
	secondary toxic (D)	8
hypothyroid	hypothyroid (E)	1
	primary hypothyroid (F)	233
	compensated hypothyroid (G)	359
	secondary hypothyroid (H)	8
binding protein:	increased binding protein (I)	346
	decreased binding protein (J)	30
general health	concurrent non-thyroidal illness (K)	436
replacement therapy:	underreplaced (M)	111
	consistent with replacement therapy (L)	115
	overreplaced (N)	110
antithyroid treatment:	antithyroid drugs (O)	14
	I131 treatment (P)	5
	surgery (Q)	14
miscellaneous:	discordant assay results (R)	196
	elevated TBG (S)	85
	elevated thyroid hormones (T)	0
no condition	(-)	6771

2.4 Model selection

1.) The project is mainly divided into two parts, the first where we predict the data on the first classifier which is the k-NN and the second part where we predict the data on the second classifier which is the Neural Network.

2.) The whole idea of the project is a multiclass classification and so initially we start off by identifying the target classes in the data and then try to perform exploratory data analysis on the data to identify the key features.

3.) After identifying them we use these features for prediction of the problem statement and then finally conclude by stating which among the two is a better classifier and which helps the endocrinologists and patients to treat and take proper medication respectively.
4.) However, we may still try to predict the problem statement with various other existing models and see how well they perform and make changes to our model to meet their performance , which could be considered as the future scope for our work.

2.5 Software packages and Tools

• Software packages: RStudio
• Libraries: ggplot2
• Development : GitHub

References

Ken Tang and Ponnuthurai N. Suganthan and Xi Yao and A. Kai Qin. Linear dimensionality reduction using relevance weighted LDA. School of Electrical and Electronic Engineering Nanyang Technological University. 2005.

Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004.

Xiaoyong Chai and Li Deng and Qiang Yang and Charles X. Ling. Test-Cost Sensitive Naive Bayes Classification. ICDM. 2004.

Vassilis Athitsos and Stan Sclaroff. Boosting Nearest Neighbor Classifiers for Multiclass Recognition. Boston University Computer Science Tech. Report No, 2004-006. 2004.

Michael L. Raymer and Travis E. Doom and Leslie A. Kuhn and William F. Punch. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 33. 2003.

Lukasz A. Kurgan and Waldemar Swiercz and Krzysztof J. Cios. Semantic Mapping of XML Tags Using Inductive Machine Learning. ICMLA. 2002.

Qiang Yang and Jing Wu. Enhancing the Effectiveness of Interactive Case-Based Reasoning with Clustering and Decision Forests. Appl. Intell, 14. 2001.

Erin L. Allwein and Robert E. Schapire and Yoram Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. ICML. 2000.

0.