International Journal of Advanced Computing Science And Engineering ISSN: 2714 - 7533
Vol. 1, No. 1, July 2019, pp. 43-47

A Study on Dengue Cases Detection based on Lazy Classifier

Nur Amiratun Nazihah Roslan [a,1,*],Hairulnizam Mahdin[b,1] , Rahmat Hidayat[2], Hendrick[3]

[1] Faculty Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[2] Department of Information Technology, Politeknik Negeri Padang, Indonesia
[3] National Kaohsiung University of Science and Technology, Taiwan
[a] nazihahroslan67@gmail.com*; [b] hairuln@uthm.edu.my;
Coresponding author*

ARTICLE INFO

Article history
Received: April 2019
Revised: June 2019
Accepted: July 2019

Keywords
Diabetes, Lazy classifier, WEKA

ABSTRACT

With the rise of social networking approach, there has been a surge of users generated content all over the world and with that in an era where technology advancement are up to the level where it could put us in a step ahead of pathogens and germination of diseases, we couldn't help but to take advantage of that advancement and provide an early precaution measures to overcome it. Twitter on the other hand are one of the social media platform that provides access to a huge data availability. To manipulate those data and transform it into an important information that could be used in many different scopes that could help improve people's lives for the better. In this paper, we gather a total of six algorithms from Lazy Classifier to compare between them on which algorithm suited the most with the diabetes dataset. This research are using WEKA as the data mining tool for data analyzation

1. Introduction

This paper is a discussion of one of the functions that are available in Weka software. Weka is one of a data mining tools with a collection of machine learning algorithms, and it is able to function for data preparation, classification, regression, clustering association rules mining, and visualization [1]. Weka is used in this research to run the algorithms and gives out its analysis on the data set that are being used. There are a total of 7 folders of classifiers that contain many more types of classifications in each one of them. In this paper we will go through all algorithms inside the Lazy classifier. In the Lazy folder, it contains 3 available algorithms which is IBK, KStar and LWL algorithm. The Lazy algorithm family will be classifying a Diabetes data set that has been provided. The dataset are being tested with different percentage splits and two different folds which are 5 and 10 folds. The results will then be presented in a table to allow the results to be previewed and analyze much more easily. This research will be able to help in analyzing and also summarizing the functions that Weka provided.

2. Research Method

This research will be focusing on exploring the "Lazy Family" Classifiers functions that are available in Weka. There are a total of 3 algorithms inside the Lazy Classification. They are IBK, KStar and LWL algorithm. They will be tested by using the diabetes data set as the subject and the result will then be presented in a table.

3. 2.1 Lazy Classifier

Lazy Learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system, as opposed to integer learning where the system tries to generalize the training data before receiving queries [2]. lazy learner simply store the data and wait until testing data appears and then conduct a classification based on the most related data within the training data that has been stored. Lazy learners have lesser training time but more in predicting time [3]

2.2 IBK

Is a instances-based learning algorithm that can learn using polynomial number of instances, a wide range of symbolic concepts and numeric functions [4].

4. 2.3 KStar

K-Means clustering is an example of unsupervised learning. It is usually used when there is unlabeled data within our dataset. For example like data without clear categories or groups. The main purpose of this algorithm is to match them into suitable groups in the data, with the number of groups presented by the variable K [5].

5. 2.4 LWL

Also known as locally weighted learning which is a class of function approximation techniques, where a prediction is done by using an approximated local model around the current point of interest

6. System Methodology

Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their database, data repositories and data warehouse. With this kind of information, companies have been able to improve their businesses by applying the patterns, relationships, and trends that have lain hidden or undiscovered within colossal amounts of data [6].

7. 3.1 Data Mining

Nowadays, with the increase in volume of medical data online which include laboratory data, represents an important and valuable resource that can provide a foundation for improved

understanding of disease presentation, response your therapy and health care delivery processes. Data mining supports these goals by providing a set of techniques built to discover similarities and relationships between data elements in large data sets. Therefore, data mining is the discovery of interesting, unexpected or valuable structures in large data sets [7].

Advances growth in Knowledge Discovery and Data Mining brings together the latest research which providing tools in statistics, databases, machine learning, and artificial intelligence that are part of the exciting and rapidly growing field of Knowledge Discovery and Data Mining The last decade has seen an explosive growth in the generation and collection of data. Advances in data collection also results in expansion use of bar codes for most commercial products, and the computerization of many business and government transactions have flooded us with data that could be generating an urgent need for new techniques and tools that can intelligently and automatically assist in transforming this data into useful knowledge

3.2 Sentiment Analysis

Sentiment analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text [8]. Usually, besides identifying the opinion, these systems extract attributes of the expressions including:

- Polarity : if the speaker express a positive or negative opinion,
- Subject : the thing that is being talked about,
- Opinion holder: the person, or entity that express the opinion.

8. Results and Discussions

Table 1: Summary results of Lazy Classifiers on Diabetes cases

	IBK					KStar					LWL				
	% SPLITS			CROSS-VALIDATION		% SPLITS			CROSS-VALIDATION		%SPLITS			CROSS-VALIDATION	
	33	66	99	5	10	33	66	99	5	10	33	66	99	5	10
Time taken to build model (sec)	0	0.01	0	0.01	0	0	0	0.01	0	0	0	0	0	0	0
Correctly classified instances (%)	69.5146	72.7969	100	70.3125	70.1823	67.9612	70.8812	87.5	69.6615	69.1406	74.7573	77.3946	75	72.3958	71.224
Incorrectly classified	30.4854	27.2031	0	29.6875	29.8177	32.0388	29.1188	12.5	30.3385	30.8594	25.2427	22.6054	25	27.6042	28.776

instances (%)															
Kappa Statistic	0.3216	0.3788	1	0.3339	0.3304	0.2829	0.297	0.7143	0.2964	0.2895	0.3907	0.3952	0.3846	0.3533	0.3469
M.A.E	0.3064	0.2729	0.0013	0.2975	0.2988	0.3331	0.3128	0.1826	0.3251	0.3275	0.3751	0.35	0.3496	0.3653	0.3684
R.M.S.E	0.55	0.5205	0.0013	0.544	0.5453	0.5072	0.4789	0.3755	0.4968	0.4969	0.4296	0.4093	0.4327	0.4379	0.441
R.A.E (%)	66.6832	60.5137	0.2839	65.4608	65.7327	72.4884	69.355	39.4918	71.5337	72.055	81.6491	77.6087	75.6372	80.3699	81.0549
R.R.S.E (%)	116.0176	111.2011	0.2707	114.1272	114.3977	106.9949	102.3036	77.4524	104.221	104.2509	90.6175	87.4354	89.2426	91.8615	92.5245
Total number of instances	515	261	8	768	768	515	261	8	768	768	515	261	8	768	768

***keywords*

2. M.A.E = Mean Absolute Error
3. R.M.S.E = Root Mean Squared Error
4. R.A.E = Relative Absolute Error
5. R.R.S.E = Root Relative Squared Error

In this table, it display the summary of the results obtained from the test run in Lazy Classification. The results are divided into different percentage splits and different cross-validation. All of the options are being put into test to compare which is the best option to be tested with the diabetes dataset.

Percentage splits are the indicator where it tells us that the 1st instances until the choosen instances are in training set and the next instances until the last are taken as the test set. This give different range towards the test. Cross validation is the indicator where the decided number of folds will invokes the learning algorithm. For example, if the number of folds are 10, Weka will invokes the learning algorithm 11 times, once for each fold of the cross validation and then a final round on the entire dataset. This will also set running performance of the algorithm, hence we can compare which algorithms are the most suitable with the diabetes dataset. The results will be further explained in the conclusion section below.

6. Conclusion

This paper has shown the analyzation of all 3 Lazy Classifier that are available in Weka, which is IBK, KStar and LWL algorithms. Based on table 1 above, it shows that IBK with a 99% percentage splits shows a much more excellent results compared to the rest of other options with Lazy Classifier. This is due to the higher percentage in correctly classifying instances and lesser percentage in errors calculation.

References

[1] "Weka 3: Machine Learning Software in Java," *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 22-Jun-2019]

[2] En.wikipedia.org. (2019). *Lazy learning*. [online] Available at: https://en.wikipedia.org/wiki/Lazy_learning [Accessed 14 Aug. 2019].

[3] Medium. (2019). *Machine Learning Classifiers*. [online] Available at: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 [Accessed 14 Aug. 2019].

[4] Aha, D., Kibler, D. and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1), pp.37-66.

[5] J. G. Cleary and L. E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure," *Machine Learning Proceedings 1995*, pp. 108–114, 1995.

[6] L. Colazzo, A. Molinari and N. Villa. "Collaborating vs. Participation: the Role of Virtual Communities in a Web 2.0 world', International Conference on Education Technology and computer, 2009, pp.321-325

[7] A. Pak and P.Paroubek "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceeding of the Seventh Conference on International Languages Resources and Evaluations, 2010, pp. 1320- 1326

[8] "Sentiment Analysis: Nearly Everything You Need to Know," *MonkeyLearn*, 20-Jun-2018. [Online]. Available: https://monkeylearn.com/sentiment-analysis/. [Accessed: 22-Jun-2019].