International Conference on Information Engineering, Management and Security
2015 [ICIEMS 2015]

ISBN	978-81-929742-7-9	**VOL**	01
Website	www.iciems.in	**eMail**	iciems@asdf.res.in
Received	10 - July - 2015	**Accepted**	31- July - 2015
Article ID	ICIEMS027	**eAID**	ICIEMS.2015.027

Disease Diagnosis using Meta-Learning Framework

Utkarsh Pathak[1], PrakhyaAgarwal[1], PoornalathaG[1]
[1]Information and Communication Technology Department
Manipal Institute of Technology, Manipal University
Manipal, Karnataka 576-104, India

Abstract: *Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy. These techniques have been very effective in designing clinical support systems because of their ability to discover hidden patterns and relationships in medical data. The main objective of this paper is to develop and implement a framework which provides considerable classification results for users who have no prior data mining knowledge. We also propose a suitable prediction model to enhance the reliability of medical examinations and treatments for diseases. We analyzed different medical records for certain disease and based on the hypothesis made on the training dataset, applied it on the test dataset and achieved disease with a good accuracy. We focus on minimizing the system dependence on user input while providing the ability of a guided search for a suitable learning algorithm through performance metrics.*

Keywords: Meta-learning framework, Dataset features, classifier.

I. INTRODUCTION

As one introduces new dataset to the system, one important step is selecting which classifier will serve with one of the best accuracies for that data. An initial assessment is time consuming since one has to decide which classifier is most suited in the given context. Thus, selecting a suitable classifier for the dataset is a complex task. Even an experienced analyst might find it very difficult to find it out. Moreover, some hidden knowledge could be present in data which adds to the problem. Here, we take up an approach which involves comparing the new problem with a set of problems for which the classifier performances are already known. First, using the meta-features that are extracted from the dataset, the dataset is plot in the space. Next, identification of the dataset which resembles the most to the new dataset is carried out using distance computation Consequently the same classifier and settings that are obtained from the near neighbour are expected to achieve similar performances on the new dataset. Thus making a structure which unites the tools important to investigate new datasets and make predictions using the learning algorithm's performance would greatly aid the novice user. This outcomes in a critical pace up and an expanded dependability on the choice of the learning algorithm. The tool we discuss is proposed in [1] and the datasets used are all of .arff format and provided by the Weka Framework. We have added a functionality of prediction; where the user uploads the test and train datasets and the prediction is done on the class attribute. The rest of the paper is organized as follows: The literature survey is discussed in section II, details regarding the proposed model is given in section III, the results obtained are given in section IV, conclusions and future scope is provided in section V, followed by the references at the end.

Cite this article as: Utkarsh Pathak, PrakhyaAgarwal, Poornalatha G. "Disease Diagnosis using Meta-Learning Framework." *International Conference on Information Engineering, Management and Security (2015)*: 168-172. Print.

II. LITERATURE SURVEY

Aha [2] proposes a system that constructs rules which describe how the performance of classification algorithms is determined by the characteristics of the dataset. Rendell et al.[3] describe a system VBMS, which predicts the algorithms that perform better for a given classification problem using the problem characteristics (number of examples and number of attributes). The main limitation of VBMS is that the training process runs every time a new classification task is presented to it, which makes it slow. The approach applied in the Consultant expert system relies heavily on a close interaction with the user. Consultant poses questions to the user and tries to determine the nature of the problem from the answers. It does not use any knowledge about the actual data. Schaffer [4] proposes a brute force method for selecting the appropriate learner: execute all available learners for the problem at hand and estimate their accuracy using cross validation. The system selects the learner that achieves the highest score. This method has a high demand of computational resources. Statlog [5] extracts several characteristics from datasets and uses them together with the performance of inducers (estimated as the predictive accuracy) on the datasets to create a meta-learning problem. It then employs machine learning techniques to derive rules that map dataset characteristics to inducer performance. The limitations of the system include the fact that it considers a limited number of data sets. Moreover, it incorporates a small set of data characteristics and uses accuracy as the sole performance measure. The use of our framework is inspired by the work done in [1] which discusses the benefits of meta-data and feature selection for mining purposes. We have used the framework as the basis of our proposed model and also added the feature to predict the diagnosis.

III. PROPOSED MODEL

In this section, we present the formal working of our framework shown in fig 1. The essential characteristic of the proposed model is

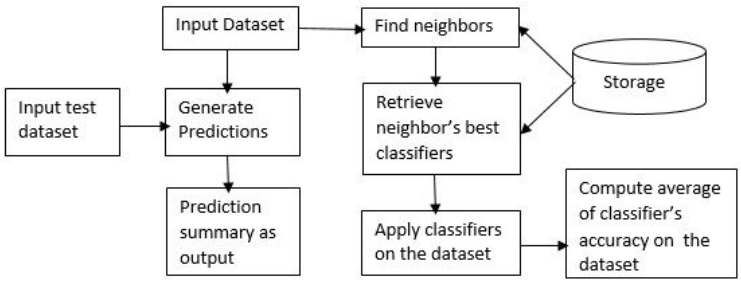

Fig. 1. Framework Model

to recommend a precise learning algorithm for a dataset submitted to the framework. The framework should achieve results with just the knowledge of the neighbour's best classifiers. The first step is to store the meta-data of the dataset. These include the total number of attributes of a dataset, the number of nominal attributes, the number of Boolean attributes and the number of continuous (numeric) attributes, the maximum number of distinct values for nominal attributes, the minimum number of distinct values for nominal attributes, the mean of distinct values for nominal attributes, the standard deviation of distinct values for nominal attributes and the mean entropy of discrete variables. Similarly for continuous attributes, it includes the mean skewness of continuous variables, which measures the asymmetry of the probability distribution, and the mean kurtosis of continuous variables representing the peak of the probability distribution. Finally, the dimensionality of the dataset is stored; It contains the overall size, represented by the number of instances, and imbalance rate information. The next step includes computing distance between the analyzed dataset and the datasets stored in the framework. The distance is computed by using the dataset metafeatures (all numeric values) as coordinates of the dataset. By representing a dataset as a point in a vector space, the distance can be evaluated using any metric defined on a vector space. The first distance computation strategy considered is the normalized Euclidean distance(E). The Euclidean distance is the ordinary distance between two points in space, as given by the Pythagorean formula. The next step is the neighbour selection step, after the distance computation phase, a list of distances is obtained, we select the Top 3 (i.e the three least distances) and we analyse the classifiers which yielded the best result on them and store the classifiers name. In the next step, we use the classifiers obtained from the last phase and use it on the analyzed dataset; we compute the average accuracy and output it to the user. Finally, we have added prediction functionality where the user uploads the test and train dataset and we predict the disease with considerable accuracy. The classifier used for prediction is J48 (studies show that J48 is more reliable than other classifiers).

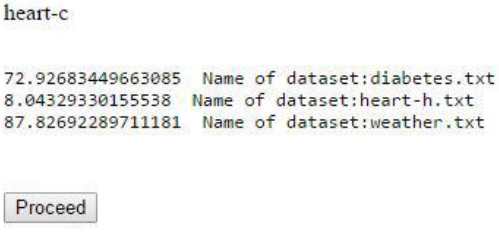

Fig. 2. Neighbours of heart-c.arff

Cite this article as: Utkarsh Pathak, PrakhyaAgarwal, Poornalatha G. "Disease Diagnosis using Meta-Learning Framework." *International Conference on Information Engineering, Management and Security (2015)*: 168-172. Print.

TABLE I. ACCURACIES IN PERCENTAGE

Dataset	J48	NaiveBayes	BayesNet	SMO	Neighbour$_a pproach$
heart-c.arff	77.55	83.49	83.49	84.15	81.73
heart-h.arff	80.95	83.67	85.03	82.65	82.08
diabetes.arff	73.82	76.3	74.34	77.34	75.17
contact-lenses.arff	83.33	70.83	70.83	70.83	75.0

IV. RESULTS

The system offers a web application which first authenticates a user; if not a valid member, a registration process is provided. Once the user is authenticated, he/she can upload a test dataset. Now, firstly the meta-features are extracted and listing of the data set features and the minimum and maximum values for each of these features is done. Next, the neighbours for the uploaded dataset is computed (i.e the top three neighbours). In the next step, the classifiers which yielded the best result on the respective neighbours is applied on the analyzed dataset and an average accuracy is computed. As mentioned earlier, the datasets used in our framework are all of (.arff) format and can be found at [6] and all the datasets mentioned at [6] are used as potential neighbours in our framework.

The Fig 1. shows neighbours results for the dataset heartc. arff ; the top three neighbour result lists out heart-h.arff ;. Now, in Table I. we have listed out the name of some of the datasets and correspondingly the accuracy (in %) obtained by the classifiers namely J48, NaiveBayes, BayesNet and SMO and the neighbour classification as the last column; In the Fig 2., we have plotted the classifier accuracy (in %) for all the datasets listed in the table. Note, our neighbour approach outperforms some of the classifiers in every dataset.

In Table II, we have computed the average classifier accuracy of every classifier over the course of all the four datasets and we find out that our neighbour approach outperforms the popular Bayesian Network Model (i.e BayesNet classifier) and performs almost as efficiently as all the othe classifiers. Now, for the prediction functionality, the user has to upload train and test datasets (see Fig 3.) and the dataset uploaded is for Prostrate Tumor and a classifier (J48) is applied on the training dataset. This step builds the decision boundary or the hypothesis model which is then applied on the test dataset (on the class atrribute) for prediction. The accuracy of prediction depends mainly on the accuracy of classification on the training dataset.

The next step comprises the display of result of the classification along with the detailed summary and the confusion matrix is presented to the user as output (shown in Fig 4.) which lists out that out of 34 samples, 9 are normal and 25 are malignant which is correct.

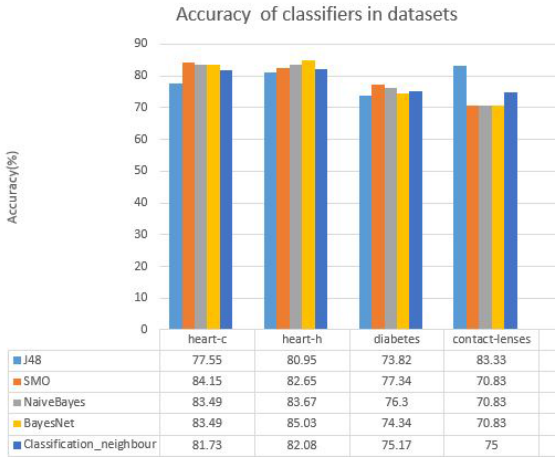

Fig. 3. Bar-graph showing accuracies of classifiers on different datasets

TABLE II. AVERAGE ACCURACY IN PERCENTAGE

J48	NaiveBayes	BayesNet	SMO	Neighbour$_a pproach$
78.91	78.57	78.42	78.74	78.50

sification along with the detailed summary and the confusion matrix is presented to the user as output (shown in Fig 4.) which lists out that out of 34 samples, 9 are normal and 25 are malignant which is correct.

V. CONCLUSIONS AND FUTURE SCOPE

The successful application of data mining in highly visible fields like e-business, stock marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare and disease prediction. In our work, we have used a framework for classification which is done by using the classifiers which yielded the best results on the neighbours of our test dataset.

Cite this article as: Utkarsh Pathak, PrakhyaAgarwal, Poornalatha G. "Disease Diagnosis using Meta-Learning Framework." *International Conference on Information Engineering, Management and Security (2015)*: 168-172. Print.

Moreover the prediction of disease functionality has also been added which makes this model highly beneficial as the complex task of predicting a disease based on patterns on similar data has been done with sufficient accuracy and diligence.

The user is presented with the option of doing classification

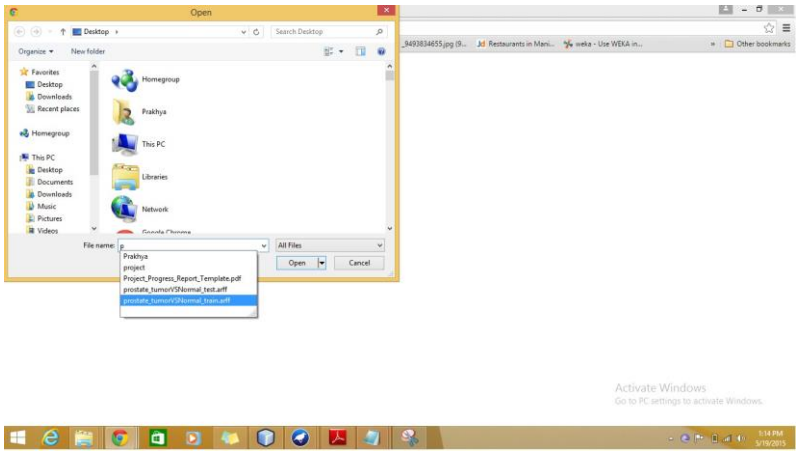

Fig. 4. Test and Train file for Prediction

```
Results
======

Correctly Classified Instances           9               26.4706 %
Incorrectly Classified Instances        25               73.5294 %
Kappa statistic                          0
K&B Relative Info Score              -1645.2619 %
K&B Information Score                  -16.4482 bits      -0.4838 bits/instance
Class complexity | order 0              33.5651 bits       0.9872 bits/instance
Class complexity | scheme            26850      bits     789.7059 bits/instance
Complexity improvement     (Sf)      -26816.4349 bits    -788.7187 bits/instance
Mean absolute error                      0.7353
Root mean squared error                  0.8575
Relative absolute error                148.4018 %
Root relative squared error            173.0394 %
Coverage of cases (0.95 level)          26.4706 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances               34

=== Confusion Matrix ===

  a  b   <-- classified as
  0 25 |  a = Tumor
  0  9 |  b = Normal
```

Fig. 5. Summary and Confusion Matrix

using the classifiers or using the neighbour approach. In our work, we found that most of the test dataset yielded a better result with the neighbour approach than the accuracy achieved by the worst classifier; thus the user can achieve healthy classification result even if he is devoid of any prior data mining knowledge.

However there is scope for further improvement; selecting the neighbours is a complex and tricky task and the number of neighbours to be found out for every test dataset is an open problem (we have taken 3 closest neighbours), the number of classifiers used could be incremented to achieve even greater accuracy.

In the prediction technique, there is a lack of extensive train and test datasets of most of the diseases as the task of accumulating the data and narrowing the number of attribute(i.e feature selection) to a limited number of attributes which affect the class attribute is a very complex task. However, the availability of the real dataset would greatly help us to learn more about disease diagnosis and prediction. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. There is a shortage of resource persons and manpower at almost every hospital, therefore an automatic medical diagnosis system would probably be exceedingly beneficial by getting positive results even from novice or inexperienced users.

ACKNOWLEDGMENT

This project consumed huge amount of work, research and dedication. Still, implementation would not have been possible if we did not have a support of many individuals and organization. Therefore we would like to extend our sincere gratitude to all of them.

Cite this article as: Utkarsh Pathak, PrakhyaAgarwal, Poornalatha G. "Disease Diagnosis using Meta-Learning Framework." *International Conference on Information Engineering, Management and Security (2015)*: 168-172. Print.

REFERENCES

[1] Potolea, Rodica and Cacoveanu, Silviu and Lemnaru, Camelia, Metalearning framework for prediction strategy evaluation, Enterprise Information Systems,Springer. page 280-295, 2011.
[2] Aha, David W, Generalizing from case studies: A case study, Proc. of the 9th International Conference on Machine Learning. Page 1-10, 1992.
[3] Rendell, Larry A and Sheshu, Raj and Tcheng, David K, Layered Concept-Learning and Dynamically Variable Bias Management, IJCAI. page 308-314, 1987.
[4] Schaffer, Cullen, Selecting a classification method by cross-validation, Machine Learning, Springer. page 135-143, 1993.
[5] Michie, Donald and Spiegelhalter, David J and Taylor, Charles C, Machine learning, neural and statistical classification, Citeseer. 1994.
[6] Sample Weka Datasets, http://storm.cis.fordham.edu/ gweiss/datamining/ datasets.html, (last accessed May 2015).
[7] Dataset Repository, http://datam.i2r.a-star.edu.sg/datasets/index.html, (last accessed May 2015).

Cite this article as: Utkarsh Pathak, PrakhyaAgarwal, Poornalatha G. "Disease Diagnosis using Meta-Learning Framework." *International Conference on Information Engineering, Management and Security (2015)*: 168-172. Print.