

Exploring Support Vector Machines and Random Forests for the Prognostic Study of an Arboviral Disease

A. Shameem Fathima

College of Computers and Information Technology,
Taif University, Saudi Arabia

L. Abdul Kadhar Sheriff

College of Administrative and Financial Sciences,
Taif University, Saudi Arabia

ABSTRACT

The need for rapid access to information in order to support critical decisions in public health cannot be disputed; however, development of such systems requires an understanding of the actual informational requirements of the practitioners. This paper explores the application of machine learning techniques for the detection of one of the Arboviral disease – Dengue. This paper reports original biological discovery through nontrivial data mining process by using accessible computational techniques. The goal of the system is to prop up the assortment, and recovery of public health documents, data, learning objects, and tools. We have deployed this standard infrastructure to facilitate data integration and knowledge sharing in the domain of dengue, which is one of the most prevalent Arboviral diseases. The proposed novel technique exhibits highly precise prediction rate (with total Mean Squared Error 0.06665807).

Keywords

Dengue fever, Data mining, Machine learning techniques, Support Vector Machine (SVM), Random Forest (RF), Feature Reduction

1. INTRODUCTION

Data mining (DM) is the intelligent computational analysis of large sets of data by using a combination of machine learning, statistical analysis and database technology, with the objective to discover patterns and rules helpful for guiding decisions about future activities [1]. Knowledge discovery in databases, or KDD for short, represents a dynamic part of research with a handful of fielded engineering applications in place. KDD means the process- starting from the collection of a mass of raw data from which we attempt to generate knowledge that should be both useful and interesting according to the objectives of the application. Current data mining tools are characterized by surplus algorithms but a lack of guidelines to select the precise method according to the nature of the problem under analysis. Learning at the base-level is focused on accumulating experience on a specific learning task e.g., medical diagnosis [2].

In medicine and medical decision making KDD and DM are critical. Neither medicine nor medical reasoning represent exact sciences, thus knowledge, which is hidden in patient records is valuable either to confirm existing theoretical, or textbook based, knowledge or to enlarge formal knowledge [3]. In medical decision support and medical reasoning, accuracy alone is not sufficient; it is however necessary to achieving relevance. The goal is to conduct research in the

identification of one of the Arbovirus - Dengue. It is of interest in learning whether or not specific data for the dengue viruses are correctly classifiable by using the concept of the machine learning namely using the SVM technique. The diagnosis and treatment of dengue is guided by the symptom and findings that the patient presents, and cannot depend on laboratory confirmation, since routine tests cannot confirm dengue with the speed required for patients in critical condition. Data mining is an analytical process that primarily involves searching throughout vast amounts of data to spot useful, but initially undiscovered patterns. The goal of data mining is prediction, generalizing a pattern to other data. The data mining process typically involves three major steps namely – exploration, model building and validation and finally deployment. The process of knowledge discovery from databases (KDD) includes several steps, such as understanding the problem domain, selecting data sources, data cleaning and pre-processing, data reduction and projection, task selection, algorithm or model selection, model evaluation and deployment [3].

A major objective of this paper is to evaluate machine learning algorithms in medical and healthcare applications to develop a model that can help make timely and accurate decisions for Arboviral diseases. In this paper, Support vector machines (SVM) classifiers and randomForest are used for analyzing the dengue data and to detect the important symptoms related to the arboviral disease. As far known, no prior comparison study was conducted between these two pioneer classification methods on a real-world information medical problem. It is shown that SVM classifiers could improve the classification accuracy significantly. The ultimate dream, of course is to have available some intelligent means that can pre-process the data, apply the appropriate mathematical, statistical and artificial intelligence techniques, and then provide a solution and an explanation. The paper is organized as follows: Section 1 provides an introduction to the topic of the research describing the problem that is discussed. Section 1.1 constitutes an overview of the various aspects of medical data mining. Section 2 describes the source of the viral data. Section 3 explains the proposed novel technique along with the limitations and constraints which the implementation imposes. In Section 4 the experimental setup is presented and section 5 the results are discussed. Finally, in Section 6 overall conclusions have been conveyed along with perspectives of future work.

1.1 Problems in Medical Data mining

Human medical data are at once the most rewarding and difficult of all biological data to mine and analyze. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database

increasingly becomes necessary. Data mining in medicine can deal with this problem. Because the medical information [4] is characteristic of redundancy, multi-attribution, incompleteness and closely related with time, medical data mining differs from other one. The major areas of heterogeneity of medical data are:

- Volume and complexity of medical data
- Physician's interpretation
- Sensitivity and specificity analysis
- Poor mathematical characterization
- Canonical form

Classification analysis is one of the widely adopted data mining techniques for healthcare applications to support medical diagnosis [5], improving quality of patient care, etc. If a training dataset contains irrelevant features (i.e., attributes), classification analysis may produce less accurate results. Feature selection is a preprocessing technique commonly used on high-dimensional data and its purposes include reducing dimensionality, removing irrelevant and redundant features [6], reducing the amount of data needed for learning, improving algorithms' predictive accuracy.

Much research work in data mining is going in improving the predictive accuracy of the classifiers by applying the techniques of machine learning such as SVM and feature selection [7]. The importance of feature selection in medical data mining is appreciable as the diagnosis of the disease could be done in this patient-care activity with minimum number of features. Feature selection [8] may provide us with the means to reduce the number of clinical measures made while still maintaining or even enhancing accuracy and reducing false negative rates. In medical diagnosis, reduction in false negative rate can, literally, be the difference between life and death.

Future developments in integrated medical data repositories, standardized data representation, and guidelines for the appropriate research use of medical is a growing volume of biomedical databases and repositories, the need to develop a set of tools to address their analysis and support knowledge discovery [5] is becoming acute. It is proposed to develop a substantial set of techniques for computational treatment of these data. The approaches in review are diverse in data mining methods and user interfaces and also demonstrate that the field and its tools are ready to be fully exploited in biomedical research.

2. SOURCES OF THE DIAGNOSTIC DATA

Professional literature broadly treats of mining medical data and proves that these techniques can be beneficial from the perspective of future diagnosis and treatment. This research is focused on developing a novel technique for predictive analysis applied to a real-life medical data set concerning the Dengue virus. The key element of a data mining study is knowing what the study is for. Conducting a survey is often a useful way of finding something out, especially when 'human factors' are under investigation. Although surveys often investigate subjective issues, a well-designed survey should produce quantitative, rather than qualitative, results [1]. That is, the results should be expressed numerically, and be capable of rigorous analysis. The first stage of the data mining process

is to select the related data from many available databases to correctly describe a given task. There are at least three issues to be considered in the data selection. The first issue is to set up a concise and clear description of the problem. The second issue would be to identify the relevant data for the problem description. The third issue is that selected variables for the relevant data should be independent of each other. [9]. The Dengue Program brought a lot of raw data in a form of surveys taken from different hospitals and diagnosis laboratories in India. Each of them was filled out by both patients and doctors participating in the screening. The data contained in these surveys contain valuable information, encapsulated in various patterns and regularities, which is used in diagnosing process in the future. The data sets used in the experiments consists of 5000 samples with 29 symptoms associated with the disease. Each sample consists of few measurements with label that denotes its symptoms. Of these, some were dengue positive and some were Dengue negative though the symptoms seemed to be like dengue positive. Using a classifier for analysis of all clinical, hematological and virological data, classification is obtained. The techniques can be used differently in different disease prevalence to yield clinically useful positive and negative predictive values.

3. THE PROPOSED NOVEL TECHNIQUE

Putting a new novel technique to work in the scientific analysis of clinical data is a demanding task. It has to overcome various practical problems, from getting access to the right data over executing the right pre- and post-processing steps to make the analysis meaningful up to problems of finding the right computational resources. The proposed technique that is constructed for an analysis platform supports the development and practical deployment of algorithms by facilitating their implementation and integration in existing workflows. The workflow of the technique is as follows:

1. Apply random Forest algorithm [10] to the entire data set and discover the vital variables that determine the symptoms highly associated with the viral disease - Dengue.
2. The features with high ($>17\%$) IncMSE from step 1 are taken as the significant variables. This value is fixed by doing the Fisher's exact Test.
3. Reduce the data set to the variables selected in step 2.
4. Apply SVM to the dataset obtained from step 3 for classifying the positive and negative dengue cases.

The following subsections describe the machine learning algorithms implemented on the data set.

3.1 Random Forest

The random Forest (RF), a classification method, is essentially a data mining package, based fundamentally on regression tree analysis and feature importance (Breiman 2001). Not only is there often a large number of records in the database, but there can also be a large number of fields (attributes, variables); so, the dimensionality of the problem is high [11]. A high-dimensional data set creates problems in terms of increasing the size of the search space for model

induction in a combinatorial explosive manner. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables. The algorithm estimates the importance [10] of a variable by looking at how much prediction error increases when data for that variable is permuted while all others are left unchanged. The arguments of the function include

| Arguments | Description |
|-----------|---|
| x | An object of class random Forest |
| type | either 1 or 2, specifying the type of importance measure (1=mean decrease in accuracy, 2=mean decrease in node impurity). |
| class | for classification problem, which class-specific measure to return. |

3.2 Support Vector Machines (SVM)

Support Vector Machines (SVMs), a global classification model [12] are supervised learning methods used for classification and regression tasks that originated from statistical learning theory. This training algorithm [12], one of the most well-known of a class of performing methods for bio-data analysis uses the idea of kernel substitution for classifying the data in a high dimensional feature space. SVMs are based on the structural risk minimization principle, closely related to regularization theory.

Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. Kernels can also be constructed to incorporate domain knowledge. This so-called 'kernel trick' gives the SVM great flexibility. The current implementation is optimized for the radial basis function kernel [13] only, which clearly might be optimal for the data. The kernel used is radial basis function and its formula:

$$\text{Radial basis function Exp } \{-\sqrt{|u-v|^2}\}$$

In the leave-one-out cross-validation, the disease status [13] of in the data set is predicted while the rest of the data is regarded as the training set. In the leave-many-out cross-validation, n individuals are uniformly at random picked up from the data set, marked and put back, where n is the size of the data set..

4. EXPERIMENTAL SETUP

In this section, the experimental setup is described; Dengue fever data is collected from different hospitals and medical diagnosis labs in India. Based on this data an information table is created from which the importance of symptoms for the dengue diagnosis is generated.

4.1 Data Set Description

Data was collected from over 5000 records of arboviral information specifically for dengue .Some were from patients who had dengue positive and others were having the symptoms of Dengue but found to be negative on doing the IgM and IgG tests. Attributes are real-valued input features of the patient record describing the symptoms such as : Age, Sex, fever, chills, Coryza , systolic, diastolic, Shock, Myalgia, Malaise, Arthralgia, Hallucinations, Confusion, Altered consciousness, Convulsion, Neck rigidity, Hemorrhagic Symptoms, Pleural Effusion, Hb, RBC*103 cells/cu.mm, WBC. These features were used to predict the disease dengue.

4.2 random Forest Usage in R

The random Forest is increasingly used statistical method for classification and regression problems introduced by Leo Breiman in 2001, to investigate two classical issues of variable selection. The first one is to find important variables for interpretation and the second one is more restrictive and try to design a good cost-conscious prediction model. The main contribution is twofold: to provide some experimental insights about the behavior of the variable importance index based on random forests and to propose a strategy involving a ranking of explanatory variables using the random forests score of importance and a stepwise ascending variable introduction strategy. Developed originally for medical applications, RF has been applied as an effective statistical tool for biological and ecological research. The random forest procedure provides two importance measures: [14]

Mean Decrease Accuracy (%IncMSE): It is constructed by permuting the values of each variable of the test set, recording the prediction and comparing it with the unpermuted test set prediction of the variable (normalised by the standard error). A higher %IncMSE value represents a higher variable importance.

Mean Decrease Gini (IncNodePurity): Measures the quality (NodePurity) of a split for every variable (node) of a tree by means of the Gini Index. A higher IncNodePurity [17] value represents a higher variable importance, i.e. nodes are much 'purer'.

4.3 SVM Usage in R

R is a language and environment for statistical computing and graphics. There are five packages that implement SVM in R: e1071, kernlab, klaR, svmPath and shogun . This work will focus on the e1071 package because it is the most intuitive. The e1071 package [15] was the first implementation of SVM in R. The *SVM()* function provides an interface to *libsvm* , complemented by visualization and tuning functions. *libsvm* is a fast and easy-to-use implementation of the most popular SVM formulation of classification (C and v), and includes the most common kernels . The R implementation is based on the S3 class mechanisms. It basically provides a training function with standard and formula interfaces, and a *predict()* method.

A viral dataset is used for dengue diagnostic to which SVM is applied. The SVM model will be able to discriminate whether the patient has dengue or not. The first function is *SVM()* [16] which is used to train a support vector machine. Some import parameters include:

data: an optional data frame containing the variables in the model.

type: sets how SVM() will work. The possible values for classification are: C, nu

kernel: defines the kernel used in training and prediction. The options are: linear, polynomial, radial basis and sigmoid

degree: parameter needed if the kernel is polynomial (default: 3);

gamma: parameter needed for all types of kernels except linear (default: 1/(data dimension));

coef0: parameter needed for polynomial and sigmoid kernels (default: 0);

The steps involved in the technique are: [17]

1. Read the dataset
2. Prepare the Dataset
3. Divide at random the dataset in two subsets, one with about 70% of the instances to training, and another with around the remaining 30% of instances to testing
4. Choose the parameters: The range to gamma parameter is between 0.000001 and 0.1. For cost parameter the range is from 0.1 until 10.
5. Train the model
6. Test the model

Both for the SVM and the partitioning tree (via rpart()), the model is made to fit and tried to predict the test set values. Predictions from the model, as well as decision values from the binary classifiers, are obtained using the predict () method.

5. COMPUTATIONAL RESULTS

In this section a dataset of dengue viral diagnostic is used and SVM is applied in it. The SVM model will be able to discriminate dengue positive and negative cases. In this dataset there are 5000 instances and 29 attributes for each instance. The data set is saved and divided at random [18] the dataset in two subsets, one with about 70% of the instances to training, and another with around the remaining 30% of instances to testing. Then choose the parameters .After training the model, run the model again the test set to predict classes. The final results for gaining features in medicine using the SVM method are tabulated for both 10 cross fold and 100 cross fold validation with cost = 100, gamma = 0.3.

Table 1. .Results of svm

| Crossfold | Total Mean Squared Error | Squared Correlation Coefficient | Number of Support Vectors |
|-----------|--------------------------|---------------------------------|---------------------------|
| 10 | 0.1008455 | 0.3102515 | 1908 |
| 100 | 0.1024797 | 0.3104216 | 1908 |

Using random Forest [19], the highly important symptoms related to the disease are analyzed and tabulated. To find the results of the package random Forest is used (formula = report ~ ., data = vdata, importance = TRUE, proximity = TRUE) and the following results are obtained

Number of trees: 500
No. of variables tried at each split: 9
Mean of squared residuals: 0.04932047
% Var explained: 60.95

Using only the important features obtained from random Forest the SVM technique is applied to the new data set and the results are analyzed. The features containing higher %IncMSE value represents the higher variable importance.

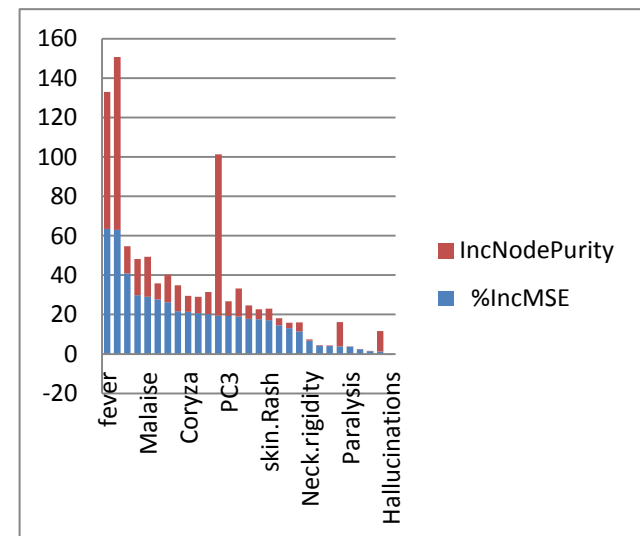


Fig 1: . Results of random Forest technique

The features with high (>17 %) IncMSE are taken as the important variables and another data set with these 18 variables as attributes is the new (Unclassified) Predictor Data. These results are compared with the real time study of the viral diseases by the doctors and virologists reported by the World Health Organization [20] .They report proves that the Patients with dengue had significantly lower platelet, white blood cell (WBC) and Signs of rash and indicators of liver damage, in combination with other variables such as age, myalgia, WBC count, and platelet counts. SVM technique is again applied to this new data and the results are tabulated as follows:

Table 2. Results of SVM after feature reduction

| Crossfold | Total Mean Squared Error | Squared Correlation Coefficient | Number of Support Vectors |
|-----------|--------------------------|---------------------------------|---------------------------|
| 10 | 0.07018363 | 0.3102515 | 1062 |
| 100 | 0.06665807 | 0.3104216 | 1062 |

6. CONCLUSION

In this paper, the motivation behind the prediction studies is discussed. The extensive computational results show great potential of the proposed prediction methods. The proposed novel technique exhibits highly precise prediction rate (with total Mean Squared Error 0.06665807). However, it is likely that infection by multiple viruses is important in development of the disease Further investigation is required to clarify correlation of viral factors with development of arboviral disease. The future work is the focus on validation with different types of arboviral diseases.

7. ACKNOWLEDGMENTS

Sincere gratitude is expressed to all the patients and primary care physicians who provided us with a cosmic amount of viral data needed for study.

8. REFERENCES

- [1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. MorganKaufmann, 2000.
- [2] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Trans. Knowledge and Data Engineering*, 8 :866-883, 1996
- [3] CIOU K.J., MOORE G.W., Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, Vol. 26, No 1-2, pp.1-24, September-October 2002
- [4] Kononenko, I. Bratko, I., and Kokar, M. (1998), Application of machine learning to medical diagnosis, in Michalski, RS, Bratko, I and Kubat M. (Eds), *Machine Learning in Data Mining: Methods and Applications*, Wiley, New York, pp. 389-428
- [5] Fernando Alonso, Juan P. Caraca – Valente and Cesar Montes, “Combining Expert Knowledge and Data Mining in a Medical diagnosis domain”, *Expert Systems with application*, pp. 367-375, Vol. 23, 2002.
- [6] Shusaku Tsumoto, Problems with Mining Medical Data, *IEEE Trans* 2000.
- [7] Ian H. W., Eibe F. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition– San Francisco: MorganKaufmann, 2005.
- [8] T. Mitchell. Machine Learning. McGraw-Hill International, 1997.
- [9] Kononenko, I. Bratko, I., and Kokar, M. (1998), Application of machine learning to medical diagnosis, in Michalski, RS, Bratko, I and Kubat M. (Eds), *Machine Learning in Data Mining: Methods and Applications*, Wiley, New York, pp. 389-428.
- [10] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [11] http://statwww.berkeley.edu/users/breiman/RandomForests/cc_home.htm
- [12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001.
- [13] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000
- [14] www.r-project.org.
- [15] cran.r-project.org/web/packages/e1071/e1071.pdf
- [16] <http://www.jstatsoft.org>
- [17] Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. A., and Tseng, T. L., “Data Mining: Medical and Engineering Case Studies” *Proceedings of the Industrial Engineering Research 2000 Conference*, Cleveland, Ohio, pp. 1-7, May 21-23, 2000.
- [18] Mertik M., Kokol P., Zalar B. Gaining Features in Medicine Using Various Data-Mining Techniques // *Computational Cybernetics ICC 2005*, IEEE international Conference. – 2005. – P. 21–24.
- [19] The R Journal Vol.2/1, June 2010
- [20] Ageep AK, Malik AA, Elkarsani MS. Clinical presentations and laboratory findings in suspected cases of dengue virus. *Saudi Med J*. 2006;27:1711–1713