Learning from Ensembles: Using Artificial Neural Network Ensemble for Medical Outcomes Prediction

Fariba Shadabi, Dharmendra Sharma and Robert Cox
School of Information Sciences and Engineering, University of Canberra, ACT, 2601, Australia

Abstract

Predicting the outcome of a medical procedure or event with high level of accuracy can be a challenging task. To answer the challenge, data mining can play a significant role. The main objective of this study is to examine the performances of an Artificially Intelligent (AI)-based data mining technique namely Artificial Neural Network Ensemble (ANNE) in prediction of medical outcomes. It also describes a novel approach, namely "RIDC-ANNE". This approach tries to improve data quality by configuring an ensemble of bagged networks as a filter and identifying the regions in the data space that have high impact on the system performance. Furthermore, it can also be used to extract explanations and knowledge from several combined neural network classifiers. The methodology employed utilizes a series of clinical datasets. The datasets embody a number of important properties, which make them a good starting point for the purpose of this research. This study reveals that the RIDC-ANNE approach can be used to successfully extract the regions in the data space that have high impact on the system performance and enhance the overall utility of current neural network models.

Introduction

Data mining techniques can be employed to support clinical data analysis. Recently AI-based data mining techniques such as Decision Trees (DT) and Artificial Neural Networks (ANNs) have drawn the attention of computer scientists and clinicians for intelligent patterns and information retrieval from clinical data sources [1], [2].

Both ANN and DT techniques have the ability to model non-linear relationships between dependent and independent variables. However research shows that with the growing power of ANN tools, ANN can often be a good analytical alternative to DT [2], [3].

Unfortunately, the success of any data mining techniques is usually dependant on the quality of the data of interest. If the data is inadequate, or contains incomplete, irrelevant and misleading information, data mining algorithms may produce less accurate and less transparent results, or may even fail to discover any useful information. Therefore, the data pre-processing stage is the most important factor that affects the success of data mining on a given task.

In earlier works, we addressed some of the practical issues associated with the use of ANN in the prediction of kidney transplant outcomes [4], [5]. In this paper we discuss our experience of applying a novel neural networks ensemble technology, known as RIDC-ANNE (Rules and Information Driven by Consistency in Artificial Neural Networks Ensemble) for the purpose of predicting the outcome of medical procedures or events. The case studies described in this paper are from "kidney transplantation", "Pima Indian diabetes" and "Wisconsin cancer" datasets. The last two datasets are stored in the UCI repository [6] and are frequently used as benchmark data.

AI-Based Data Mining Tools in Clinical Problem Domains

AI-based data mining tools have been used successfully in a number of medical domains. A good example is the famous Pima Indians diabetes dataset. This dataset is based on personal data of the Pima Indians, which originally obtained from the US National Institute of Diabetes, Digestive and Kidney Diseases. The dataset is now available in the UCI repository [6]. All patients in this dataset are females and at least 21 years old of Pima Indian heritage. The database contains 768 data samples taken from patients who may show signs of diabetes. All samples in this dataset have no missing attributes. There are 8 attributes (inputs) in this database and two output classes, diabetes and non-diabetes. The attributes are age, pregnancy information and medical measurements. Over the years there have been a huge amount of researches related to this database. Interestingly a comparison study conducted by Abczewski and Duch in 2001, reported only 77% accuracy rate for the best classification performance [7].

1-4244-0674-9/06/$20.00 ©2006 IEEE.

The Wisconsin Breast Cancer dataset is another popular data set that has been used as benchmarks by many researchers [6]-[9]. This is a relatively clean and non-complex dataset. The dataset was originally obtained from the University of Wisconsin Hospitals, in Madison from Dr. William H. Wolberg. The data set has nine attributes (inputs) and two output classes. All nine inputs are continues and range from 1 to 10. The database contains 699 samples with 683 complete data and 16 samples with missing attributes. Each of the 683 available instances is labeled as either Benign *(444 instances or 65% of data) or* Malignant *(35%). The task is to predict benign or malignant classes. Classification accuracy of this database is generally approximated to 90% or higher. The same comparison study reported by Abczewski and Duch in 2001 indicated 97% accuracy for the best classification performance [7].*

Recently, there has been substantial research in predicting graft outcomes and detecting key parameters influencing graft outcome [10]-[12]. The factors that will determine the outcome of graft transplants is still unclear, although certain factors such as age of patient, number of Human Leukocyte Antigen (HLA) mismatches have been known to have an influence on the outcome of transplants. For this study, we use neural networks ensemble to try to predict the outcome of kidney transplants by using a small trial data set made available to us from a database of kidney transplantations [5], [13]. For the purpose of this study some variables were removed because they were actually an indication of the outcome of the transplant. The variables that were retained are AGE (Recipient age at transplant), MISA (Number of mismatches A), MISB (Number of mismatches B), MISDR (Number of mismatches DR), MISDQ (Number of mismatches DQ), REFHOSP (Referring hospital), REFSTAT (Referring state), DONHOSP (Donor hospital), DONSTAT (Donor state), TRANHOS (Transplant hospital), TRANSTA (Transplant state), DONSOUR (Donor source), DONAGE (Donor age), DONSEX (Donor sex), ISCHEMIA (Total ischemia to nearest hour), and KIDPRESI (Initial kidney preservation). In this scenario the challenge is to select the right kidney from the available pool of organs for a particular patient, thereby maximizing the chances for the successful transplantation. It should be noted that this dataset is greatly corrupted with missing features, and random errors in the values of the features. We especially chose this dataset since it stands as a good example of clinical data with complex characteristics.

The following section presents our methodology presented in this study in which we investigated the use of the RIDC-ANNE technique for the prediction of outcomes of above medical events.

Methodology

The RIDC-ANNE method is motivated by the Rule Extraction approach of [14]-[16]. In this strategy the ensemble of networks computes the target concept and the input vectors are the actual network's input vectors. This strategy also tries to consider the diversity and expertise of the component networks in the rule generation process. It should be mentioned that unlike previous method in [16], the RIDC-ANNE method does not focus on identifying the ensemble members that are relevant in explaining the prediction (output) associated with a particular "case". Instead it tries to explain the output of an ensemble based on a "cluster of cases" that consistently generates agreement (on their class labels) across the classifiers with similar expertise. This form of regularization could offer a significant improvement in comprehensibility of predictions and help the user to investigate different regions of data space.

Overall, for the purpose of this study, the following methodology was employed:

1. Pre-process the data set. This includes: extracting the data from different tables, cleaning the data, transforming the nominal attributes into numeric attributes, choosing the appropriate parameters to be included in the dataset with the help of domain expert and normalization.
2. Split the dataset to three equal parts for training, overtraining prevention (tuning) and testing (mainly with balanced distribution of success and failure cases).
3. Perform classification using a series (n=500 or n=100) of Multilayer Perceptron (MLP) network that were trained independently, to differentiate between two classes.
4. Generate new training sets by extracting the patterns (examples) that were consistently causing x% agreement (for x= 50 To 100) across the ANN classifiers, in the testing phase.
5. From the new training sets generated in step 4, choose a reasonably big training set that has provided both a good level of accuracy (based on its corresponding classification table) and a reasonable amount of model agreement.
6. Modify the new training set (selected from step 5) by replacing the desired classes of all cases with their corresponding class labels (i.e. the class assigned by the trained ensemble).
7. Grow a decision tree [17] from the selected samples generated in step 6.
8. Analyze the results.

Results for Kidney Transplantation

In this section we study closely the results of the kidney transplantation data. Using the above methodology, the balanced test set reached 70% accuracy rate with 87%-agreement among the networks (435 of 500 networks), based on 19% of data points. The results are shown in Fig. 1. It should be mentioned that the accuracy rate also reached 76% with 89%-agreement among the networks (446 of 500 networks), using only 63 examples (14% of data points).

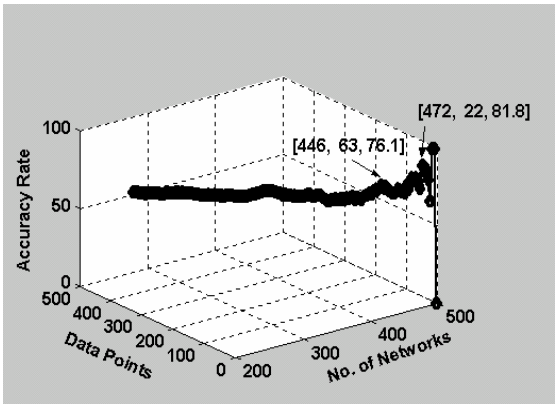

Figure 1. The results for neural network ensemble with 500 bagging (classifiers).

For rule generation stage using the decision tree tool, at first, we applied a more conventional strategy and generated a new training data set by feeding the entire examples in the test set to a trained ensemble and replacing the true class labels of the original test instances with the class labels assigned to them by the ensemble. The rule set generated by decision tree tool produced 26 rules.

Furthermore, in order to study different regions of the data space and also make the rule set substantially easier to understand, we enforced the model to mainly consider the examples whose class labels consistently caused agreement across the ANN classifiers. In effect, this strategy tries to remove some branches of rules and identify the regions that have strong impacts on the system performance (this also means identifying the data spaces that have been consistently misclassified). The following rule set was produced by applying RIDC-ANNE approach based on the 84 examples whose class labels (outputs from the ensemble) were in agreement across 87% of classifiers:

1. *donstate <= 4 AND donage <= 43:* ***Success***
2. *donhosp <= 109 AND misa <= 1 AND refstate > 3 AND misb > 0 AND refhosp > 94:* ***Failure***
3. *misa <= 1 AND misb <= 1:* ***Success***
4. *donsex > 0 (i.e. Female):* ***Failure***
5. *misa > 1:* ***Failure***
6. *age > 27:* ***Success***
7. *Else:* ***Failure***

As it can be seen, in this experiment, the rule set produced fewer rules (only 7 rules). These rules were valid for 97 % of cases (i.e. 82 cases) in the data set.

Results for UCI datasets

The investigation of the kidney data described above was also applied to two datasets, namely "Pima Indian diabetes" and "The Wisconsin cancer".

In this experiment, for the Pima Indian diabetes data, using the majority voting the balanced test set reached slightly above 77% accuracy rate (very similar to previous studies). However by using only 93% (i.e. 239 cases) of test data points, it was possible to achieve 80% accuracy rate with 68%-agreement among the networks (68 of 100 networks). The results are shown in Fig. 2. It should be mentioned that the accuracy rate also reached 87% with 99%-agreement among the networks, using only 152 examples (59% of data points).

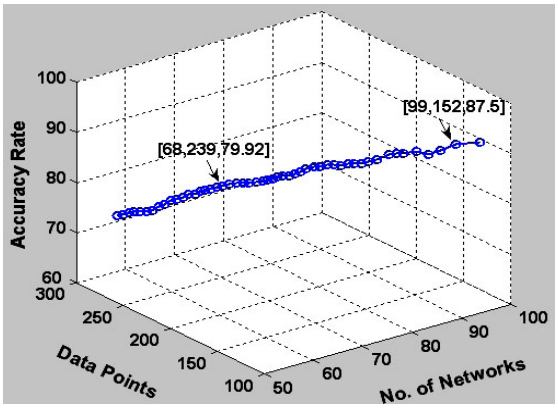

Figure 2. The results for the Pima Indian diabetes data with 100 bagging.

Furthermore, the RIDC-ANNE approach can be used to generate a set of rules that are roughly used by ANNE to predict the diabetes and non-diabetes cases. The following rule set was produced by applying RIDC-ANNE approach for classifying the "diabetes" and "nondiabetes" samples, using the 239 examples that their class labels (outputs from the ensemble) were in agreement across 68% of classifiers:

1. *2-hour OGTT plasma glucose < 139.5 AND diabetes pedigree function < 0.76:* ***Diabetes***

2. 2-hour OGTT plasma glucose < 139.5 AND diabetes pedigree function > 0.76 AND Body Mass Index < 36.5: ***Diabetes***

3. 2-hour OGTT plasma glucose < 139.5 AND diabetes pedigree function > 0.76 AND Body Mass Index >36.5: ***Non-Diabetes***

This means, for this selected data set, the first two rule sets are roughly used by ANNE to predict the diabetes cases with accuracy rate of 94%. However it should be noted that for the "healthy" (non-diabetes) cases presented in the above samples the ANNE consistently misclassified the samples (i.e. only 48% accuracy rate). In this case 68% of classifiers made the same mistake and reported 52% of healthy cases as diabetes cases. An domain expert can use these information and rules to highlight and study the regions in the data space that have negative impacts on the generalization ability of classifiers and more suitable input vectors may be supplemented accordingly.

Not surprisingly, for the Wisconsin cancer dataset, using the majority voting the balanced test set reached 97.8% accuracy rate. This model was able to classify about 98% of cancer cases and 97% of non-cancer cases. The accuracy rate also reached 98.2% with 65%-agreement among the networks (68 of 100 networks), based on 99.5% (i.e. 227 cases) of test data points. The results also revealed that it is possible to achieve 98.6% accuracy rate with 100%-agreement among the networks based on 96.5% (i.e. 220 cases) of test data points. In other words, by removing only 0.5% (i.e. 8 cases) of test data points, it was possible to generate 100%-agreement among the networks and a higher level of accuracy. The complete results for the Wisconsin cancer dataset are shown in Fig. 3.

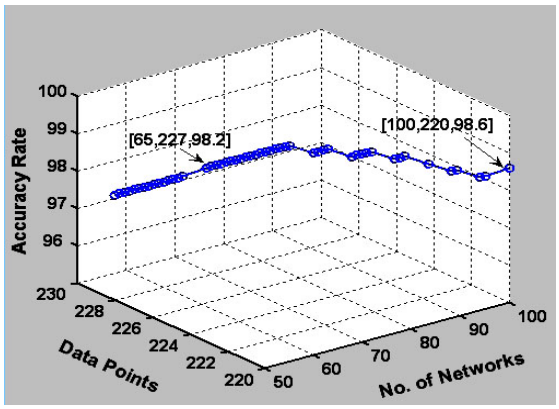

Figure 3. The results for the Wisconsin cancer dataset with 100 bagging.

The following rule set was produced by applying RIDC-ANNE approach using the 220 examples whose class labels (outputs from the ensemble) were in agreement across 100% of classifiers:

1. Uniformity of Cell Shape < 3.5 AND Bland Chromatin > 4.5: ***Cancer***

2. Uniformity of Cell Shape < 3.5 AND Bland Chromatin < 4.5: ***Non-Cancer***

3. Uniformity of Cell Shape > 3.5: ***Cancer***

Conclusion

Over the past few years there has been great interest in the development of robust and efficient data mining algorithms and a variety of approaches have been proposed. These newer algorithms and approaches have provided great opportunities for researchers to carry out different clinical trials and enhance current prediction techniques. Unfortunately, their success is usually dependant on the quality of the data of interest. This study described a novel approach, namely "RIDC-ANNE" that is designed to improve data quality by configuring an ensemble of bagged networks as a filter and identifying the regions in the data space that have high impact on the system performance

This approach facilitates data mining by combining the power of 'black box' connectionist learning systems such as ANNs with 'transparent' rule based decision-making methodologies. The primary experimental results reveal that the RIDC-ANNE approach can be used to identify the regions in the data space that have strong impacts on the system performance. Furthermore, it can also be used to translate a neural network ensemble into an alternative more understandable model. Overall, a system user or a domain expert may be able to use the extracted information and rules to:

a) Identify unrealistic predictions and perhaps pin points the source of theses sort of predictions by studying different regions of data space and their corresponding rule sets. In other words, the system can be used to identify data spaces that have been consistently misclassified. This can also highlight the regions in the data space that have negative impacts on the generalization ability of neural networks and more suitable input vectors may be supplemented accordingly.

b) Predict in which section of input vectors or under which circumstances the neural network model may perform poorly.

It is important to mention that for the purpose of this study we have examined a series of medical datasets and compared our results with the results obtained by other methods whenever possible. However for the rule extraction part, in general, different structure of rules may be extracted for different data partitions of a problem domain [7]. Therefore it should be noted that even for the same problem domain, it is impossible to make a good comparison of the rules that have been generated for different regions of the data space in this study with the rules obtained by different methods by other researchers.

References

[1] Sheppard, D., McPhee, D., Darke, C., Shrethra, B., Moore, R., Jurewitz, A., Gray, A. "Predicting Cytomegalovirus disease after renal transplantation: an artificial neural network approach." *International Journal of Medical Informatics*, Vol. 54, 55-76, 1999.

[2] Michie, D., Spiegelhalter, D.J, Taylor, C.: "Machine Learning, neural nets and statistical classification." Ellis-Horwood, Chichester, 1994.

[3] Astion, M.L. and Wilding, P. "The application of backpropagation neural networks to problems in pathology and laboratory medicine. " *Arch. Pathol. Lab. Med*. Vol. 116 995–1001. 1992.

[4] Shadabi, F., Cox, R., Sharma, D., Petrovsky, N. "Use of Artificial Neural Networks in the Prediction of Kidney Transplant Outcomes." *KES'04, the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems* , Vol 3. pp. 566-572, 2004.

[5] Shadabi, F., Cox, R., Sharma, D., Petrovsky, N. "Experiments with A Neural Network Ensemble to Predict Renal Transplantation Outcomes. "*AISAT 04 Proceeding, the 2nd International Conference On Artificial Intelligence Science and Technology*, pp. 271-276, November 2004.

[6] C.J. Mertz, P.M. Murphy, UCI repository of machine learning databases, available at the address: http://www.ics.uci.edu/pub/machine-learning-databases

[7] Duch, W. Adamczak, R. and Grabczewski, K. "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules." *IEEE Transactions on Neural Networks,* Vol 12, pp. 277-306, 2001.

[8] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", *SIAM News*, Vol. 23, No. 5, pp 1 & 18. September 1990.

[9] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", *Proceedings of the National Academy of Sciences*, U.S.A., Vol. 87, pp 9193-9196. December 1990.

[10] Rapaport F.T.: "The current status of the HLA controversy in clinical transplantation." *Transplant. Proc*. Vol.27, No. 1, 1995.

[11] Doyle H., et al. "Predicting outcomes after liver transplantation. A connectionist approach." *Ann. Surg*. Vol. 219, No. 4, pp.408-415, , 1994.

[12] Matis, S., Doyle, H., Marino, I., Murad, R, Uberbacher, E.: "Use of Neural Networks for Prediction of Graft Failure following Liver Transplantation." *Proceeding of the Eight Annual IEEE Symposium on Computer-Based Medical Systems*, 1995.

[13] ANZDATA, Data Dictionary: ANZDATA Registry Database, 2000.

[14] Zhou, Z. H. and Jiang, Y., "Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble" *IEEE Transactions on Information Technology in Biomedicine* Vol.7, No.1, pp. 37-42, 2003

[15] Zhou, Z. H. and Jiang, Y., "NeC4.5: neural ensemble based C4.5." *IEEE Transactions on Knowledge and Data Engineering* Vol. 16, No. 6, pp. 770-73, 2004.

[16] Wall, R., Cunningham, P., Walsh, P., and Byrne, S., "Explaining the output of ensembles in medical decision support on a case by case basis", *Artificial Intelligence in Medicine* Vol 28, pp. 191-206, 2003

[17] Quinlan, J. "Bagging, boosting, and C4.5." *Proceedings of the Thirteenth National Conference on Artificial Intelliegnce***,** pp.725-30, 1996.