Oriental Journal of Computer Science & Technology Vol. **3**(2), 311-316 (2010)

Multiple classifiers system for medical diagnosis

M. SOLOMON PUSHPARAJ[1] and P.J. KULKARNI[2]

[1]MCA Department, Sinhgad Institute of Management and Computer Application, Pune (India).
[2]Department of Computer Science and Engineering, Walchand College of Engineering, Sangli (India).

(Received: July 29, 2010; Accepted: September 03, 2010)

ABSTRACT

Data mining helps in decision making. Due to the peculiar feature of the medical profession, physician desperately needs a helping tool to take an efficient and intelligent decision. Good performance, the ability to appropriately deal with missing data and with noisy data (errors in data), the transparency of diagnostic knowledge, the ability to explain decisions, and the ability of the algorithm to reduce the number of tests necessary to obtain reliable diagnosis are the various features desired from the machine learning classifier to solve the medical diagnostic task. Every machine learning method has its own features and no single method can provide all the desired features. We solved this problem by using multiple machine learning methods. In this paper we developed multiple classifiers system which helps the physician in the time of decision making process. Backpropagation algorithm (ANN), K-NN Algorithm (CBR) and Modified towing splitting rule algorithm (CT) are used in this system. We tested the system with three different disease datasets like diabetes, heart disease, breast cancer. It showed better results in reliability and performance which two are most desired features in the medical diagnostic task.

Key words: Data mining, artificial neural network, case based reasoning, classification tree, medical diagnosis.

INTRODUCTION

Data mining is the nontrivial extraction of implicit, previously unknown, interesting and potentially useful information from data[1]. Data mining helps in decision making. Now a day's hospitals and health care institutions are well equipped with monitoring and other data collection devices. Data is collected and shared with other hospital information systems. Due to the idiosyncracies of the medical profession, physician desperately needs a helping hand to take an efficient and intelligent decision.

Good performance on diagnostic accuracy, the ability to appropriately deal with missing and with noisy data, the transparency of diagnostic knowledge, the ability to explain decisions are the some of the desired features expected from the good machine learning system for medical diagnostic tasks[2]. Lot of machine learning algorithms are available (like backpropagation in neural network, K-NN from Case based Reasoning and modified towing splitting rule in Classification Tree) in the market, but the main problem is not a single machine learning algorithm has all the expected features for the medical diagnosis tasks. For example backprobagation algorithm has very good diagnostic accuracy performance but it has poor transparency and explanation ability. K-NN has the explanation ability but it does not have generalization. It's transparency of knowledge representation is poor. CT model has very good transparency but it has very ordinary performance to handle the missing and noisy data. Every single algorithm has its own merits and demerits. Physician needs a new approach which will have more supporting features to the medical diagnosis task.

Present Study

´ We have created a model called multiple classifiers system using multiple machine learning algorithms which has more expected features compared to a single machine

learning algorithm (Backprobagation, K-Nearest Neighbor and modified towing splitting rule) for medical diagnostic tasks.

- We have tested the proposed model with three different disease datasets like diabetes, heart disease, breast cancer and found its improved performance.
- We have implemented the model using XLMiner software which made the system very easy to use.

Related Literature

Classification, clustering, prediction, association, rule extraction and sequence detection are the various types of problems we can solve through data mining. The techniques used in data mining are from different fields like statistics, machine learning and pattern recognition. Machine Learning is the study of computer algorithms that improve machine learning automatically through experience[3]. Abdel and Kenneth summarized the principle of machine learning approaches to ECG classification. They evaluated and proved that machine learning algorithms are highly accurate in medical diagnosis[4]. Artificial neural network, Case based reasoning and Classification Tree algorithms are coming under the machine learning field.

Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science[5]. Applications include bankruptcy prediction[6], handwriting recognition[7], speech recognition[8], product inspection[9], fault detection[10], medical diagnosis[11 - 13] and bond rating[14].

Case-based reasoning (CBR) is an approach to problem solving that emphasizes the role of prior experience during future problem solving[15]. CASEY gives a diagnosis for the heart disorders[16]. GS.52 is a diagnostic support system for dysmorphic syndromes. NIMON is a renal function monitoring system, COSYL that gives a consultation for a liver transplanted patient[17] and ICONS presents suitable calculated antibiotics therapy advised for intensive care patients[18]. Medical Informatics Research Group at Ain Shams University developed successful applications in cancer and heart diseases[19].

Tzung and Gang applied decision tree methods to medical data mining problems [20]. Christine and Hamish developed a medical diagnosis system using classification tree (FT Tree) and an LR model (FT LR). It predicted the probability of a patient with chest pain is having an myocardial infarction (MI)[21].

Some researchers used Hybrid approach in the medical diagnosis task. Siddharth and Shruthi developed the design of a two tier Neural Internetwork based Medical Diagnosis System (NIMD) that uses k-Nearest Neighbor Classification for Diagnosis pruning. The system is essentially two tiered with the first tier handling diagnosis pruning. The second tier consists of separate modules for each disease that handles the actual detection of the disease based on the intensities of the various symptoms reported by the patient[22]. David and Magnus designed a Decision Support System for Parkinson's disease. They proposed a method based on ANNs and SVMs to aid the physician in the diagnosis of PD[23].

Many systems have been developed based on single ANN, CBR and CT methods. We proposed multiple classifiers system using the above three methods for medical diagnostic task.

Proposed Model

Every single machine learning algorithm has its own advantages and disadvantages. Backpropagation algorithm (ANN), K-NN algorithm (CBR) and Modified towing splitting rule (CT) are used in the new proposed model. The working procedure of multiple classifiers system is explained below in the form of an algorithm and flowchart in detail.

Algorithm

Step-1

Create a two different Datasets namely S1 and S2 for training and testing the multiple classifiers system.

Step-2

Train the multiple classifiers system (which has a separate ANN, CBR and CT classifier inside) using the Training Dataset S1.

Step-3

For all the input data in the testing dataset

S2 do the following steps.

- Calculate the outputs using the ANN and CBR classifiers only.
- Compare the outputs. If they are same then ANN and CBR classifiers output will be the output of the multiple classifiers system. Otherwise once again pass the input test data to the CT classifier. The CT Classifier's output will be the output of the multiple classifiers system.

EXPERIMENTAL

We tested the proposed multiple classifiers system with 3 different diseases datasets like diabetes, heart disease, breast cancer. Table 1 shows misclassification performance of individual Artificial Neural Network, Case Based Reasoning and Classification Tree models and the proposed multiple classifiers system for the Pima Indian diabetes disease dataset. The diabetes dataset has taken from the URL. http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. From the overall 532 cases 319 cases are used for training and remaining 213 test cases are used for testing the classifier performance.

Table 2 shows misclassification performance of individual Artificial Neural Network, Case Based Reasoning and Classification Tree

Table 1: Performance table for diabetes disease dataset

Iteration Number	Misclassification Using ANN	Misclassification Using CBR	Misclassification Using CT	Misclassification Using Proposed Model
1	57	57	55	54
2	59	43	47	49
3	64	54	54	53
4	49	44	51	43
5	48	43	54	48

Table 2: Performance table for heart disease dataset

Iteration Number	Misclassification Using ANN	Misclassification Using CBR	Misclassification Using CT	Misclassification Using Proposed Model
1	22	14	31	19
2	18	16	16	15
3	16	14	27	15
4	26	24	25	22
5	22	24	28	23

Table 3: Performance table for breast cancer disease dataset

Iteration Number	Misclassification Using ANN	Misclassification Using CBR	Misclassification Using CT	Misclassification Using Proposed Model
1	11	9	13	9
2	6	7	14	7
3	15	3	11	7
4	8	5	16	8
5	14	10	18	11

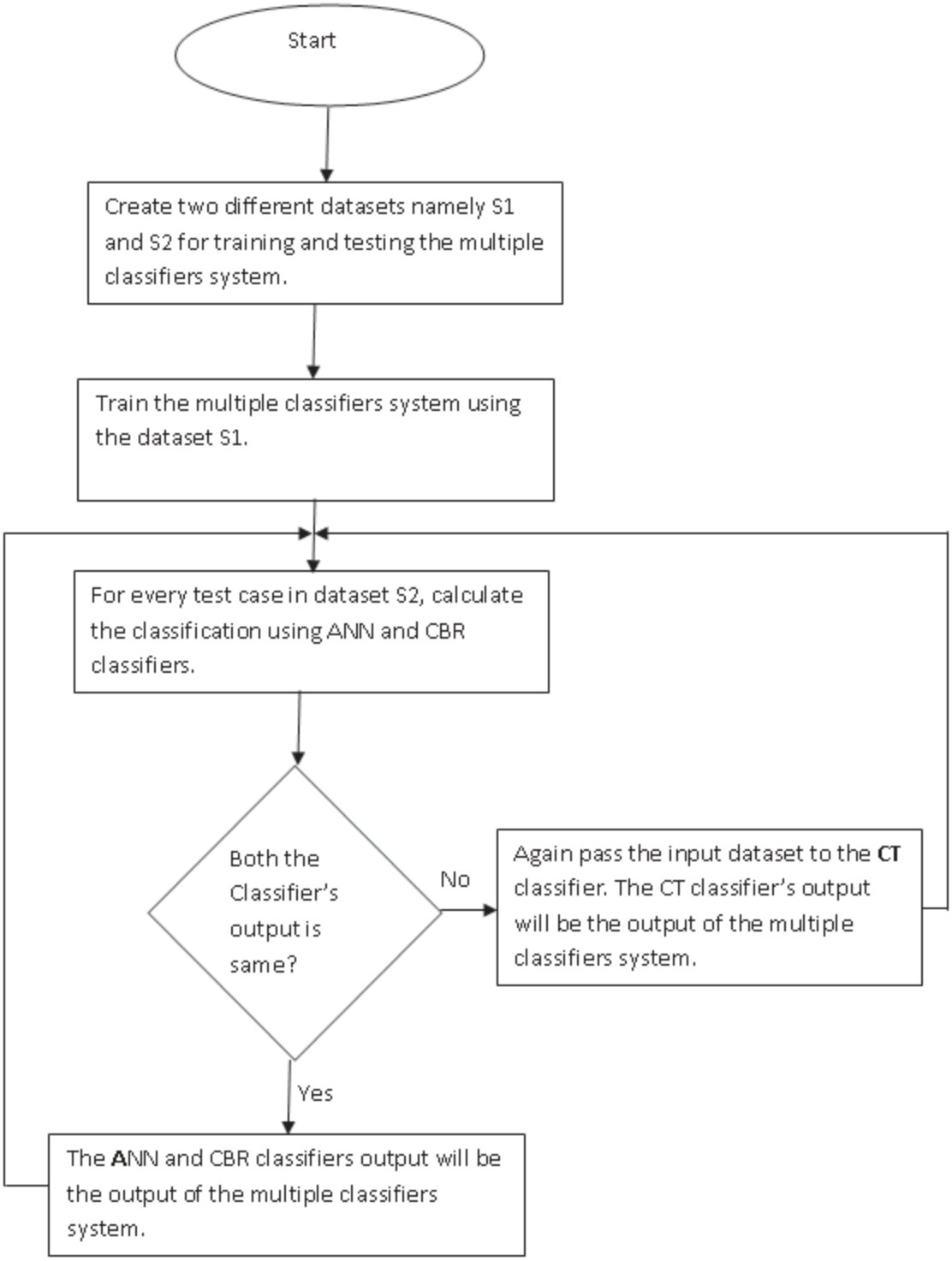

Fig. 1: Multiple Classifiers System

models and the proposed multiple classifier system for the Heart Disease dataset. The Heart Disease dataset has taken from the UCI machine learning dataset URL http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. From the overall 290 cases 174 cases are used for training and remaining 116 test cases are used for testing the classifier performance.

Table 3 shows misclassification performance of individual Artificial Neural Network, Case Based Reasoning and Classification Tree models and the proposed multiple classifiers system for the Breast Cancer Wisconsin (Original) Data Set. The Breast Cancer dataset has taken from the UCI machine learning dataset http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original). From the overall 643 cases 386 cases are used for training and remaining 257 test cases are used for testing the classifier performance.

CONCLUSION

The performance of the classifier depends on the dataset it is used for training and testing. Every machine learning algorithm has its own merits and demerits. There is no single machine learning algorithm which is going to give the best result for all the type of datasets. Data mining in medical field is a challenging task because of the complexity in the medical domain. In this research multiple classifiers system gave the reliability (i.e. the result is going to be supported by more than one algorithm) and performance which is the two top most expected priorities in the medical diagnosis task.

REFERENCES

1. W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 213-228 (1992).
2. Kononenko. I, Machine learning for medical diagnosis: History, state of the art and perspective.
3. Tom Mitchell, Machine Learning, McGraw Hill (1997).
4. Abdel-Badeeh M. Salem, Kenneth Revett, El-Sayed A. El-Dahshan, Machine Learning in Electrocardiogram Diagnosis, Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 429 – 433 ISSN 1896-7094 ISBN 978-83-60810-22-4.
5. B.Widrow, D. E. Rumelhard, and M. A. Lehr, Neural networks: Applications in industry, business and science, Commun. ACM, **37**: 93-105 (1994).
6. E. I. Altman, G. Marco, and F. Varetto, Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience), *J. Bank. Finance*, **18**: 505-529 (1994).
7. I. Guyon, Applications of neural networks to character recognition, *Int. J. Pattern Recognit. Artif. Intell.* 5: 353-382 (1991).
8. H. Bourlard and N. Morgan, Continuous speech recognition by connectionist statistical methods, IEEE Trans. Neural Networks, **4**: 893-909 (1993).
9. J. Lampinen, S. Smolander, and M.Korhonen, Wood surface inspection system based on generic visual features, Industrial Applications of Neural Networks, F. F. Soulie and P. Gallinari, Eds, Singapore: World Scientific, 35-42 (1998).
10. E. B. Barlett and R. E. Uhrig, Nuclear power plant status diagnostics using artificial neural networks, Nucl. Technol., **97**: 272-281 (1992).
11. W. G. Baxt, Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion,Neural Comput., **2**: 480-489 (1990).
12. H. B. Burke, Artificial neural networks for cancer research: Outcome prediction, *Sem. Surg. Oncol.*, **10**: 73-79 (1994).
13. H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R. Marks, D. P. Winchester, and D. G. Bostwick, Artificial neural networks improve the accuracy of cancer survival prediction,

Cancer, **79**: 857-862 (1997).

14. S. Dutta and S. Shekhar, Bond rating: A nonconservative application of neural networks, in Proc. IEEE Int. Conf. Neural Networks, vol. 2, San Diego, CA 443-450 (1988).
15. R. Mantarasi, D. Mcsherry, Retrieval, reuse, revision and retention in case-based reasoning, The Knowledge Engineering Review, **20**(3): 215–240, Cambridge University Press (2006).
16. Kolodner, J. Case-Based Reasoning, Morgan Kaufmann, San Mateo (1993).
17. M. Lenz, S Wess, H Burkhard and B Bartsch, Case based reasoning technology: from foundations to applications, Springer (1998).
18. B. Heindl. Et al,: A Case-Based Consiliarius for Therapy Recommendation (ICONS) computer-based advise for calculated antibiotic therapy in intensive care medicine, computer methods and programs in biomedicine **52**: 117-127 (1997).
19. Abdel-Badeeh M. Salem, Case Based Reasoning Technology for Medical Diagnosis, Proceedings of world academy of science and technology. **25:** (2007).
20. Tzung-I Tang·Gang Zheng·Yalou Huang·Guangfu Shu·Pengtao Wang, A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis, IEMS Vol. 4, No. 1, pp. 102-108, (2005).
21. Christine L. Tsien, Hamish S. F. Fraser, Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction, MEDINFO 98,B. Cesnik et al. (Eds), Amsterdam: IOS Press.
22. J. B. Siddharth Jonathan and K.N. Shruthi , A Two tire neural inter network based approach to medical diagnosis using K-nearest neighbor classification for diagnosis pruning.
23. David Gil A, Magnus Johnson B, “Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines”, Global Journal of Computer Science and Technology, 63-70.