Classification Rule Discovery for Diabetes Patients by using Genetic Programming

Raj Kumar, Rajesh Verma

Abstract*—The learning algorithms have great application in knowledge discovery. Learning algorithm offers new beneficiary ways in for real-world applications. Genetic Programming (GP) have some advantages due to which it become suitable for classification in data mining for Knowledge Discovery(KDD) This paper focuses on the classification by using the Genetic Programming. There are various types of the traditional classification techniques like Naïve Bayesian, ID3, C4.5, CART, kNN, k-mean, SVM etc. The proposed algorithm is implemented on the Diabetes data set and excremental results are compared with traditional approach

***Index Terms*— Classification, DM, GP, KDD,**

I. INTRODUCTION

Data mining or knowledge discovery is needed to make sense and use of data. Knowledge discovery in the data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of data [1] Data mining is the core step of knowledge discovery in database (KDD) and interdisciplinary field includes database management system, machine learning, statistics, neural network, fuzzy logic etc. Any of the technique may be integrated depending on the kind of data to be mined. The research in KDD is expected to generate a large variety of systems because diversity of disciplines to be contributed. Therefore a comprehensive classification system is required able to distinguish between the systems and identify the most required by the user. The major issue involved with the classification rule mining is to identify a dataset for a small number of rules to serve as classifier for predicting the class of any new instance. The classification algorithm should be accurate, simple and efficient. The existing classification algorithm assuming that the input data is drawn from a pre-defined distribution having stationary majors. Therefore these algorithms perform poorly when used to infer real world datasets. [2].

This paper presents a new genetic algorithm to improve the accuracy of the classification.

II. GENETIC ALGORITHM FOR CLASSIFICAITON

In this paper we proposed GA utilization in the classification process. It is done by random selection of given population to produce the new generation by applying the genetic operators. A binary function is used as fitness function. An abstract tree is used in the representation of selection objects; it will reduce the negative aspects of Genetic operators. The initial generation of programs uses the GA as shown below:-

1. Start
2. Specify the individual length(l)
3. GENERATION=1
4. WHILE not terminate DO
5. FOR GEN= 1 to l
   Generate random GEN_j
6. Store the individual created in the initial population
7. End Do
8. Exit

The step-1 of the algorithm is to indicate the initialization of the classification process. Step-2 defines the length of Genome. Step-3 indicates the first generation. Step-4 defines the loop till generation exists. . Step-5 generates the random gene with individual chromosome using the genetic operator. Step-6 stores the individual created in the initial population. Step-7 indicates that termination criterion is fulfilled.

III. IMPLEMENTATION AND EVALUATION OF PROPOSED ALGORITHM

3.1 Diabetes Patients Data Set:

The proposed algorithm is applied on the data of the diabetes patients [3]. The given dataset of the diabetes patients has 768 instances with the following nine attributes

1. Pregnancies(PRG): Number of pregnancies
2. PG Concentration (PGC): Plasma glucose at 2 hours in an oral glucose tolerance test.
3. Diastolic BP (DBP): Diastolic Blood Pressure (mm Hg).
4. Tri Fold Thick(TFT):Triceps Skin Fold Thickness (mm)
5. Serum Ins(SI): 2-Hour Serum Insulin (mu U/ml)
6. BMI: Body Mass Index: (weight in kg/ (height in m)^2)
7. DP Function(DPF): Diabetes Pedigree Function
8. Age: Age (years)
9. Diabetes(DBT): Whether or not the person has diabetes

However in the diabetes patients data set there are 768 instances are there but in this paper we have used only 20 instances , the dataset used for the evaluation is given in the below table 1.

prg.	pgc	dbp	tft	si	bmi	dfp	age	dbt
6	148	72	35	0	33.6	0.627	50	0
1	85	66	29	1	26.6	0.351	31	1
8	183	64	0	0	28.3	0.672	32	0
1	89	66	23	94	28.1	0.167	21	1
0	137	40	35	168	43.1	2.288	33	0
5	116	74	0	0	25.6	0.201	30	1
3	78	50	32	88	31	0.248	26	0
10	115	0	0	0	35.3	0.134	29	1

Manuscript received September 02, 2012.

Raj Kumar, Assitant Professor, Deptt. of Computer Sc. & Engg., Jind Institute of Engg. & Technology,Jind (Haryana), India,).

Dr. Rajesh Verma, Professor & Head, Deptt. of CSE, Kurukshetra Institute of Technology & Management, Kurukshetra (Haryana), India.

2	197	70	45	543	30.5	0.158	53	0
8	125	96	0	0	0	0.232	54	0
4	110	92	0	0	37.6	0.191	30	1
10	168	74	0	0	38	0.537	34	0
10	139	80	0	0	27.1	1.441	57	1
1	189	60	23	846	30.1	0.398	59	0
5	166	72	19	175	25.8	0.587	51	0
7	100	0	0	0	30	0.484	32	0
0	118	84	47	230	45.8	0.551	31	0
7	107	74	0	0	29.6	0.254	31	0
1	103	30	38	83	43.3	0.183	33	1
1	115	70	30	96	34.6	0.529	32	0

Table1: Data Set

In the above table for the attribute Diabetes(dbt) binary values are used, binary "1" represents the healthy i.e. the persons having no diabetes and the binry "0" represents the sick i.e. the person having diabetes. This data set is divided in the following two data sets.

(i) **Training Data Set:**
Training data set is the subset of data set which is used to make the model

(ii) **Applied data set:**
The applied data set is the whole data set which is applied on the model tested by the training data set.

3.2 Confusion Matrix:

A confusion matrix is accuracy measurement tool for data mining classification. Accuracy of a classifier on a given test set is the percentage of test tuples that are correctly classified by a classifier[5]. A confusion matrix is generally of the form:

		Actual Class	
		Class 1	Class 0
Predicted Class	Class 1	TP	FN
	Class 0	FP	TN

Table 2: layout of confusion matrix

Where

TP: True positive
FP: False Positive
FN: False Negative
TN: True negative

$$Accuracy = \frac{\text{Nmber of tuples classified currently}}{\text{Total number of tuples}}$$

i.e

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.3 Accuracy measurement

The above data set is implemented on the Dicipulus 5.1, which is a Genetic Programming(GP) tool for binary classification. and Notitia software is used for the data cleaning and transformation, which the inbuilt software of the Discipulus.

3.3.1 Accuracy measurement for Training data set

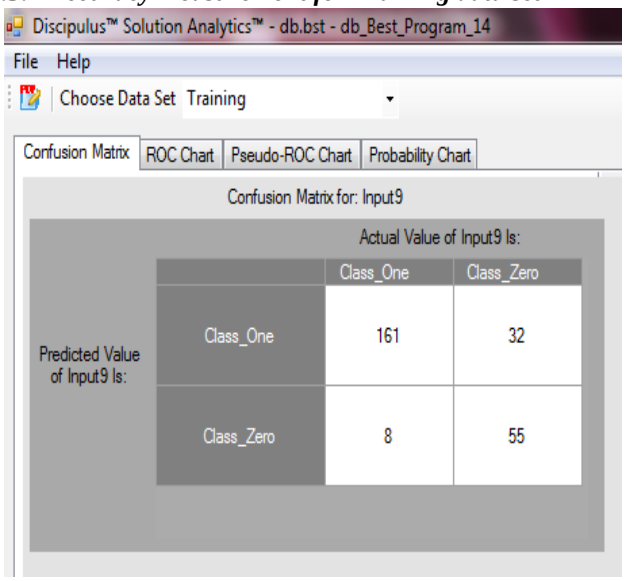

Fig. 1: Confusion Matrix for Training Data Set

Where

True Positive (TP) =161
False Positive (FP) =8
False Negative (FN) =32
True Negative (TN) =55

In the above confusion matrix for training data set 161 members of Class1 are predicted correctly. 32 members of Class 0 are predicted as class 1. 55 member of class 0 are predicted correctly and 8 members of Class 1 are predicted as Class 0.

$$Accuracy = \frac{161 + 55}{161 + 55 + 32 + 8}$$

$$= 0.84375$$

So we can say that accuracy for the classification of training data set is 84.38%. So the proposed algorithm is able to get a prediction error about 16.62% for the training data set.

3.3.2 Accuracy measurement for Applied Data Set

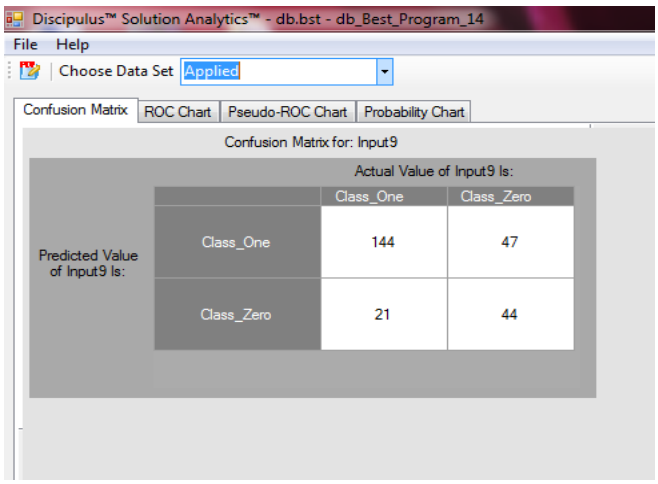

Fig. 2: Confusion Matrix for Applied Data Set

Where

True Positive (TP) =144

False Positive (FP) =21

False Negative (FN) =47

True Negative (TN) =44

In the above confusion matrix for training data set 144 members of Class1 are predicted correctly. 47 members of Class 1 are predicted as Class 1.44 member of class 0 are predicted correctly and 21 members of Class 1 are predicted as Class 0.

$$Accuracy = \frac{144 + 44}{144 + 44 + 47 + 21}$$

$$= 0.72868$$

So we can say that accuracy for the classification of Applied Data Set is 72.87%. So the proposed algorithm is able to get a prediction error about 27.13% for the Applied Data Set.

3.4 Comparative Analysis

To compare the performance of the Genetic Programming with the traditional classification methods we compare the accuracy of the proposed algorithm with the C4.5 [4]. The accuracy results are given in the below table 3.

	Proposed GP Algorithm		C4.5 Algorithm	
	Training	Test (applied)	Training	Test
Accuracy	84.38%	72.87%	57.9%	52.4%

Table3: Accuracy Comparision

The C4.5 gives prediction error about 42.1% and 47.6% for the training and test data respectively. However the proposed algorithm gives prediction error about 16.62% and 27.13% for the training and test data respectively.

IV. CONCLUSION AND FUTURE WORK

So from the implementation of the proposed algorithm and comparison with the traditional algorithm, it can obsereved that the performance of classification using the Genetic Programming is high than that of the traditional classification techniques. This technique can be applied to many other fields like financial fraud detection, whether forecasting and to detect the natural disasters like earth quake, tsunami, cloud bursting etc.

REFERENCES

1. Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy", Advances in Knowledge Discovery and Data Mining", (Chapter 1), AAAI/MIT Press 1996.
2. Raj Kumar, Dr. Anil Kr. Kapil, Anupam Bhatia ,"Modified Tree Classifiation in Data Mining Global Journal of Computer Science and Technology, Vol. 12, Issue 2 (Ver. 1.0), 2012, pp. 59-62
3. http://www.liacc.up.pt/ML/statlog/datasets/diabetes/diabetes.doc.html
4. R.Quinlan, "C.5: Programs for Machine learning," Morgan Kaufimann, 1993
5. Jaiwei Han, Micheline Kkamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006, pp 360-361

AUTHORS PROFILE

Raj Ku;mar is working as Asst. Prof. in the deptt. of Computer Sc. & Engg. at Jind Institute of Engg. & Technology, Jind(Haryana), India. He obtained his M.Tech. degree form Chaoudary Devi Lal University, Sirar and MCA degrees from Krurukshetra University, Kurukshetra. His area of Research is Data Mining

.

Dr. R.K Verma is working as Professor and Head in Deptt. of CSE at Harayna Institute of Technology & Management, Kurukshetra(Haryana), India. He obtained his Ph.D. in Computer Sc. and MCA From Kurukshetra University, Kurukshetra (India), M.Tech in CSE from KSOU, Karnatka (India). His Research area includes: Software Engg., Mobile Ad-Hoc Networks, Web Technology, Data Mining and Knowledge Management