International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 3, Issue 4, 2012, pp541-545
http://bipublication.com

APPLICATION OF FUZZY ID3 TO PREDICT DIABETES

Daveedu Raju Adidela[#1], Lavanya Devi. G[2], Jaya Suma. G[3], Appa Rao Allam[4]

[1,2]Department of Computer Science and Systems Engineering, College of Engineering,
Andhra University, Visakhapatnam, Andhra Pradesh, India
[#]123.davidjoy@gmail.com
[3]Department of Computer Science and Engineering
GITAM University, Visakhapatnam, Andhra Pradesh, India.
[4] Director, CRRao AIMSCS, UoH, Hyderabad, India

[Received-11/09/2012, Accepted-26/11/2012]

ABSTRACT-

Diabetes is one of the challenging diseases that the human race is facing which was not newly invented but prevailing from ages. Many of the scientists around the globe have been working on diabetes and making aware of the symptoms and effects of the disease on the various organs of the body. This paper proposes a Hybrid Classification System (HCS) from which the occurrence of diabetes can be predicted. The sample data which influence the causing of diabetes is supplied to this system as input. The system adopts three phases. In the first phase, clustering of the data using EM-algorithm is performed. The second phase carries out the classification of the obtained individual clusters using fuzzy ID3 (Iterative Dichotomiser 3). As of the second phase of the process, adaptation rules are obtained. These rules are essential in the prediction of diabetes. In the third phase the test tuple is supplied to the rules to predict the class label.

Key words- *adaptation rules; clustering: E-M; diabetes; fuzzy ID3; k-fold cross- validation.*

I. INTRODUCTION

The importance of the disease was revealed by discovering its causes by French and Greek physicians. Dr. Basting and his team invented the functionality of the insulin that plays the vital role in causing the diabetes. According to the inventions, Diabetes exist in two types, i.e. Type-I (Insulin Dependent) and Type-II (Non-Insulin Dependent) diabetes and it is exposed in the year in 1959. According to WHO, around 300 million people would be affected by this decease by the year 2025 around the world. It is observed that the most of the people affected by Type-II (90-95%) diabetes and it is genetic in nature. There are several risk factors that develop Type- II diabetes; some of them are hyper tension, obesity, over consuming of alcohol and fat diet, desk-bound nature, aging and many more.

Traditionally physicians will go through an array of tests, which include the above said and also specific organic tests [1][2][3]. There are other diagnostic methods [4] which are under practice. But with the advent of computational technology in data mining, the researchers are concentrating on diagnosis and analysis tools for predicting the diseases which are challenging the mankind.

II. LITERATURE SURVEY

There are several clustering algorithms available to mine the data and these are used in several disciplines

[5][6][7]. Clustering techniques are further classified depending upon the methodologies that are being used. They are named from hierarchical method to co-clustering techniques for high dimensional data [7][8]. Similarly, the classification techniques also play a vital role in analyzing the data and to predict the information [9]. Some of the techniques used for predicting the diseases by using classification technique are Genetic Algorithms, Neural Networks, Fuzzy Logic, Case Based Reasoning (CBR), Bayesian Classification and Support Vector Machine (SVM). They are being used depending upon the problem specificity. These techniques have their own drawbacks and advantages.

- Neural Networks [10]: It has high accuracy but the process is highly complex.
- CBR [11]: It become more intelligent, with the increase in number of cases in data base and it is intuitive. But there is a difficulty of finding good similarity metric.
- Genetic Algorithms [12]: Used in dynamic applications and convergence is granted, but time to convergence is uncertain. It has inherent parallelism but the fitness function is crucial for the development of a suitable solution of a problem.
- Fuzzy System [13]: It allows dealing with vagueness or inexact data but proficiency is needed.
- SVM [14]: It can classify both linear and non-linear data but high complexity is involved in the process.
- Bayesian Systems [15]: It predicts class membership probabilities but it does not hold the class conditional independence.

III. DATA PREPROCESSING

There is a possibility of non existing, in consistent and noise data that is extracted from data base, data warehouse, World Wide Web and other repositories which are multiple and heterogeneous data. This raw data should be preprocessed to get the accurate and consistent data and to eliminate accumulation of mistakes and increase the efficiency of the further processing. There are several phases in the pre processing and they are well known with data cleaning, data integration and transformation, and data reduction.

3.1. Data cleaning

In this phase the data is to be cleaned by placing the appropriate data in missing values, the outlier (the data that is not comply with the existing data) data to be identified and removed and the noise data (confusing data) to be polished. Data is polished by using binning (referring neighborhood and distributing into bins) and regression methods. The outliers are eliminated by clustering methods.

3.2. Data integration and transformation

Data integration binds the data from multiple sources into a consistent data store. The schema integration is performed by knowing Meta data of the relevant attribute. The redundancies are ascertained by using the correlation methods.

Data transformation transforms the data into suitable form for mining. It involves smoothing, aggregation which is needed at the time of summary, generalization i.e. from low level data to higher level data, normalization which is an array of attribute data between two limits and attributes construction to smoothen the mining process.

3.3. Data reduction

In data reduction process, the huge data is reduced to a limited data that facilitates the mining without deviating from integrity.

By using above methodologies, the quality of the data is improved and the refined data is then forwarded to the HCS system as an input and after processing the data adaptation rules are obtained.

3.4. Nature of sample data

There are several factors which influence the cause of diabetes like obesity, over consumption of alcohol and fat foods, hyper tension, aging and physical inactivity. There are other specific factors that imply the cause of diabetes such as in take of certain medicines, infected pancreas or damaged pancreas which fails to produce insulin properly and gestational diabetes (caused during pregnancy).

Occurrence of the gestational diabetes is more prone for women who conceive after 25 years, diabetic affected family history, over weight before pregnancy and giving birth to a baby whose weight is more than 4 kilograms. The number of times pregnancy is also predominant factor for causing diabetes. The commonly observable facts are taken for consideration for predicting the occurrence of diabetes. It is named here as tested positive. The data comprises of nine different attributes including one action attribute elaborated below with their short forms:

1. Number of times pregnant- preg
2. Plasma glucose concentration after 2 hours in an oral glucose tolerance test- plas
3. Diastolic blood pressure (mm Hg)- pres
4. Triceps skin fold thickness (mm) - skin
5. 2 - Hour serum insulin ([mu] U/ml) - insulin
6. Body mass index (weight in kg/(height in m)^2) -mass
7. Diabetes pedigree function - pedi
8. Age (years) - age
9. Class variable - class

The last action attribute or class attribute has two classes, i.e., tested_positive and tested_negative for diabetes. Sample data is given below

Table : sample diabetes-patients data

preg	plas	pres	skin	Insu	mass	pedi	age	class
2	88	58	26	16	28.4	0.766	22	N*
1	119	86	39	220	45.6	0.808	29	P*
8	120	80	48	200	38.9	1.162	41	N
1	89	66	23	94	28.1	0.167	21	N
0	137	40	35	168	43.1	2.288	33	P
1	79	80	25	37	25.4	0.583	22	N
3	78	50	32	88	31	0.248	26	P
10	79	60	42	48	43.5	0.678	23	N
2	197	70	45	543	30.5	0.158	53	P
8	100	74	40	215	39.4	0.661	43	P
4	164	82	43	67	32.8	0.341	50	N

*P - tested_positive, N - tested_ negative

IV. OVERVIEW OF THE SYSTEM

The system has two phases of processing the data. In the first phase, the cleaned data is given to the EM (Expectation and Maximization) clustering algorithm, which clusters or groups the data into several clusters depending upon the similarity of the objects by using expectation and maximization steps [16]. The precision of the cluster is verified by using various statistical approaches like the gap statistic [17], [18]. The genuine number of clusters are identified and used for further processing which determines the effectiveness of classification.

In the second phase each cluster is separately processed by the fuzzy ID3 algorithm and a decision tree is produced for each cluster. This implies production of adaptation rules from individual decision trees. Ninety percent of the data is given to train the system; remaining ten percent is used for testing the system. K-fold cross validation is applied to the system for evaluating the accuracy of the classifier.

The data is divided into k-mutually exclusive subsets D_1, D_2,……,D_K each with equal size. For the iteration 'i', the partition D_i is reserved for test data and remaining subsets used to train the model. The iteration is preformed for k-times. The total number of correct classifications from k-iterations divided by the total number of iterations gives the accuracy. And also confusion matrix for positive and negative tuples is obtained to measure the accuracy of the classifier.

In the third phase, the test data which doesn't have any class label is reasoned with the adaptation rules and class label is extracted. The tuple with obtained class label will forecast the cause of diabetes.

V. CLUSTERING CASE LIBRARY USING E-M ALGORITHM

This section attempts to partition the data base into several clusters using E-M (Expectation-Maximization) algorithm [16]. EM method is simple and easy to implement and it is an iterative method for finding maximum likelihood where the model depends on unobserved latent variables.

E-M algorithm assigns each object to a cluster according to a weight representing the probability of membership.

The EM Algorithm is elaborated below.

Step 1: This step calculates the probability of cluster membership of case x_i, for each of the cluster.

Each case x_i is assigned to cluster C_k with the probability,

$$P(x_i \in C_k) = P(C_k/x_i) = P(x_i).P(x_i/C_k)/P(x_i)$$

Where, $P(x_i /C_k) = N(m_k, E_k(x_i))$ follows the normal distribution around mean(m_k) with expectation (E_k).

Step2: This maximization step computes the distribution parameters and their likelihood given the data.

The probability estimate from the above is used to estimate the model parameters, for example

$$m_k = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i.P(x_i \in C_k)}{\sum_j P(x_i \in C_j)}$$

This step is the maximization of the likelihood of the distributions given the data.

EM iteration alternates between step1 and step2.

To view the clusters formed, calculate the cluster - maximum and cluster - minimum for each cluster from the calculated mean and standard deviation of each cluster by using above algorithm.

VI. OVERVIEW OF FUZZY ID3 ALGORITHM

Fuzzy ID3 decision technique [19] is the extension of ID3 (The Interactive Dichotomizer 3) technique [20]. ID3 is most widely used decision tree algorithm and proposed by Quinlan. The fuzzy ID3 is applied to fuzzy set of data (data with membership grades) to generate a fuzzy decision tree. The ID3 uses maximum information gain or decrease in entropy of an attribute to decide the splitting attribute of the tree. But fuzzy ID3 uses membership function of the fuzzy sets for the attributes. The data is divided in to fuzzy subsets (sometimes by the opinion of the data expert) by using different graphical representation (Gaussian, Triangle etc) to determine the membership function of the individual attributes. The information gain is calculated for all the attributes and selects an attribute which has maximum information gain for the root node. Divide the data into fuzzy subsets and find the new membership values of these sub nodes which is the product of the membership value of the root node and the membership value of the fuzzy sets of the other respective attributes. Again the gain is calculated for the test attribute and repeat the process until the leaf node is achieved. As the data fuzzification is used for the attribute entropy and information gain formulation, the formulas slightly varies from ID3. Every produced node is pruned to reduce the effect of noise and exceptions. This pruning uses the threshold value θ_n defined on the minimum number of patterns and other threshold value θ_r defined on the proportion of the class values. These two values are set to get the optimal values depending on problem specification. To be a leaf node the number of patterns of a node should less than θ_n and proportion of a class is greater than or equal to $\theta_{r.}$

VII. CONCLUSION

The HCS system is useful for predicting the disease from huge sample data set as it initially clusters the data and applies the classification on clusters. Whereas the traditional classification methods when applied on the whole data leads to complexity and also poor in accuracy.

This paper proposes a hybrid classification system which builds from EM algorithm for clustering and fuzzy ID3 algorithm to obtain decision tree for individual clusters. The decision tree of each cluster produces a set of adaptation rules from which diabetic effected patient is identified. The eight attributes of the individual person is given to the system and by utilizing these adaptation rules, the system will predict the person as tested_positve or tested_negitive of diabetes. The accuracy of the classification is measured by confusion matrix and k-fold cross-validation.

VIII. FUTURE SCOPE

This proposed model is worthy to pay attention as it further can be refined by including the attributes, which are sensitive to causing of diabetes like diabetic history of the family, eye vision, food habits and other parameters like gender, social support, habits etc [21]. This model can be extended for the prediction of other diseases like hyper tension, acidity, ulcers, retinopathy, kidney disorders, heart diseases and cancer. This model can also be successfully implemented in forecasting the weather and rain falls, agricultural yields, business transactions, economical crisis and speculation business in stock market.

REFERENCES

[1] http://www.med.umich.edu/mdrtc/cores/chemcore/hemoa1c.html

[2] http://www.diabetes-and-health.com/diagnostic-tools.html

[3] http://diabetes.webmd.com/guide/diabetes_diagnosis_tests

[4] Markus W Büchler, Marc E Martignoni, Helmut Friess and Peter Malfertheiner, "*A Proposal for a New Clinical Classification of Chronic Pancreatitis*" BMC Gastroenterology, 2009.

[5] Arabie, P., and Hubert L.J, "*An Overview of Combinatorial Data Analysis*," In P Arabie, L.J Hubert, G.D Soete (eds.), Clustering and Classification, World Scientific, River Edge, NJ, pp. 5–63, 1996.

[6] Massart, D. and Kaufman, L., "*The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*,". John Wiley & Sons, New York, NY, 1983.

[7] Pavel Berkhin., "*Survey of Clustering Data Mining Techniques,*" Technical Report, Accrue Software, San Jose, CA, 2002.

[8] Anderberg, M., "*Cluster Analysis and Applications,*" Academic Press, New York.(co-clustering), 1973.

[9] Witten, I., Frank, E., "*Data Mining: Practical Machine Learning Tools and Techniques,*" 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[10] Widrow, B., Rumelhart, D.E., Lehr, M.A., "*Neural Networks: Applications in Industry, Business, and Science,*" Communications of the ACM, vol. 37, pp. 93-105, 1994.

[11] CBR features available at http://www.ai-cbr.org

[12] Mitchell Melanie., "*An Introduction to Genetic Algorithms,*" MIT Press, Cambridge, MA, 1996.

[13] Yi-lai Chen, Tao Wang, Ben-sheng Wang and Zhou- jun Li, "*A Survey of Fuzzy Decision Tree Classifier,*" Fuzzy Information and Engineering, vol. 1, no. 2, pp. 149-159, 2009.

[14] Shigeo Abe, "*Support Vector Machines for Pattern Classification*", Springer, 2nd edition, 58-59, 2005.

[15] Kazuo J. Ezawa and Til Schuermann, "*A Bayesian Network Based Learning System, Architecture and Performance Comparison with Other Methods,*" in Symbolic and Quantitative Approaches to Reasoning and Uncertainty: Lecture Notes in Artificial Intelligence 946, C. Froideveaux and J. Kohlas, eds., Springer-Verlag, Berlin, pp. 197-206, 1995.

[16] Dempster, A.P., Laird, N.M., and Rubin, D.B., "*Maximum Likelihood From Incomplete Data Via The E-M Algorithm,*" Journal of .Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[17] Robert Tibshirani, Guenther Walther, Trevor Hastie, "*Estimating the Number of Clusters in a Data Set Via the Gap Statistic,*" Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.

[18] Gordon, A.D., "*Identifying Genuine Clusters In A Classification,*" Journal of Computational Statistics & Data Analysis archive, vol. 18, no. 5, pp 561 – 581, 1994.

[19] Umanol, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., Kinoshita, J., "*Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems,*" Proceedings of the third IEEE Conference on Fuzzy Systems, vol. 3, pp. 2113-2118, 1994.

[20] Quinlan, J.R., "*Induction on Decision Tree,*" Machine Learning, vol. 1, pp. 81-106, 1986.

[21] Sridhar, G.R. et al, "*Living With Diabetes: Indian Experience,*" Diabetes and Metabolic Syndrome: Clinical Research and Reviews, vol. 1, no. 3, pp.181-187,200