I J C T A, 9(26) 2016, pp. 385-388

Thyroid Disease Analysis and Accuracy Prediction Using Data Mining Techniques

K. Rajam[1] and R. Jemina Priyadarsini[2]

ABSTRACT

Thyroid nodule is one of the indicative of thyroid cancer .nodule can be due to the growth of thyroid cells or a cyst in the thyroid gland.so diagnosing thyroid disorder disease is a high interest to data miners, and decision trees have been useful data mining tools to diagnose the disease, but the accuracy of decision trees has been limited due to insufficient data. Recent studies demonstrated that thyroid nodules can be found in about 66%of the adult population.in order to generate more accurate decision trees for liver disorder disease this paper suggests a method based on over-sampling in minor classes to compensate the insufficiency of data effectively .experiments were done with representative algorithms of decision trees, CART, and a data set, UCI machine repository for thyroid disease and showed the validity of the method.

Keywords: Sampling, thyroid disorder disease, classification.

1. INTRODUCTION

Thyroid disease diagnosis is one of the very difficult and deadly tasks, because it needs lots of experience and knowledge. Thyroid is one of the largest endocrine gland. It is a small butterfly shaped gland which is located. The traditional ways for diagnosis thyroid disease is doctor's examination or a number of blood tests. The thyroid releases two principal hormones. The first is called thyroxin (T4). another is triiodothyronine (T3) in blood stream. Data mining is a process of analysing large data sets to find some patterns. These patterns can be helpful for prediction modelling. Data mining plays a vital role in medical field for disease diagnosis these data provide a basis for the analysis of risk factors for many diseases.

Decision trees learning uses a decision trees as a predictive model which maps observations about the items target value.it is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees.in these tree structures, leaves represent class labels and branches that lead to those class labels. A Tree can be learned by splitting the source set into subsets based on an attribute value test.

Another issue is random sampling. Sample is a subset of individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stages during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of *k* individuals.

This process and technique is known as simple random sampling, and should not be confused with systematic random sampling. A simple random sample is an unbiased surveying technique. In section 2, we provide the related work to our research, and in sections 3 we present our method of experimentation. Experiments were run to see the effect of the method in section 4 finally section 5 provides conclusions and future work

[1] Research Scholar, Computer Science, Bishop Heber College, Tiruchirapalli, Tamilnadu, India

[2] Assistant Professor, Computer Science, Bishop Heber College, Tiruchirapalli, Tamilnadu, India

2. RELATED WORK

Deepika koundal [1] have provided the information about the existing automatic tools which are available to formulate the existing automatic tools which are available to formulate the disease diagnosis part easier with efficient way. Also different and trends are also investigated.

Edgar Gabriel [2] have proposed two parallel versions of code that are used for texture-based segmentation of thyroid FNAC images which is a critical first step in realizing a fully automated CAD solution.an MPL version of the code is developed to exploit distributed memory compute resources such as PC clusters

Nikita Singh [3] has been proposed classification using SVM, KNN and Bayesian. Also the information about segmentation and classification methods which are very important for medical image processing is also provided efficiently. The results shows that SVM gives better accuracy as compared to KNN and Bayesian.

Estuations G [4]have suggested a computer-aided diagnosis(CAD)system prototype named as TND(Thyroid Nodule Detector).it is used for the detection of nodular tissue in ultrasound(US)thyroid images and videos acquired during thyroid US examinations.

Won –jin moon [5] have done in her paper the evaluation on the diagnostic accuracy of ultrasonography (US) criteria for the depiction of benign and malignant thyroid nodules.it is done by using issue diagnosis as the reference standard. they concluded that shape, margin, echogenicity and presence of calcification are important criteria for the discrimination of malignant from benign nodules.

3. THE METHOD OF EXPERIMENTATION

We are interested in finding better decision trees for UCI Thyroid disorder data set because the data set is relatively small and has somewhat high error rate, we what to compensate the property of disdaining minor classes in splitting branches, it is highly possible that instances of minor classes are treated in the lower part of the tree, and this treatment may increase misclassification rate for minor classes.so we want decision tree algorithms to treat the instances of minor classes more importantly.in order to do this we increase the number of instances of minor classes by duplication. The following is a brief description of the procedure of the method.

INPUT: Thyroid disorder data set.

OUTPUT: Decision trees.

Begin

Do random sampling of size of 180, nine times.

For each sample data set Do

Generate a decision tree for the sample data:

Do While the accuracy of decision tree increases:

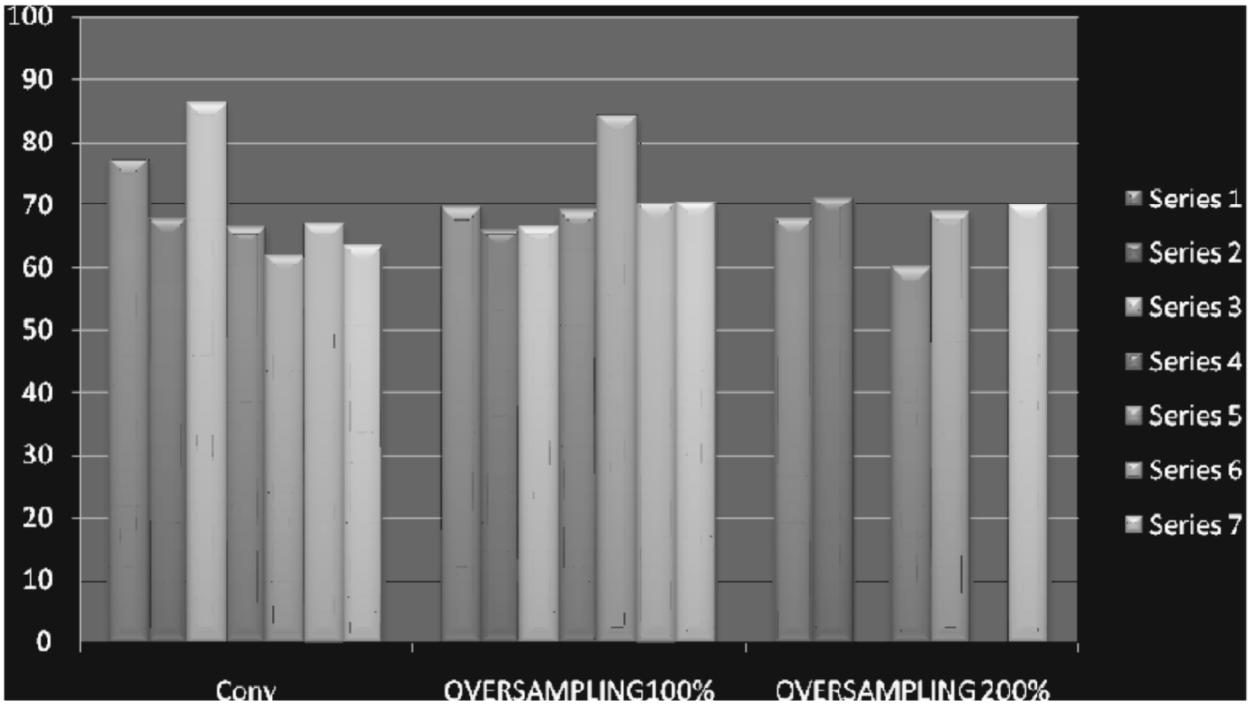
Duplicate the instances of minor class:

Generate a decision tree:

End while:

End Do;

End.



In the algorithm we duplicate the instances of minor class by 100 percents until the accuracy of generated decision tree decreases.the sample size is half of the data set so that we have large enough data set for testing

4. EXPERIMENTATION

Experiments are run using a UCI machine learning repository data set (26)called thyroid Disorder (1) to watch the effect of the method. The number of instances is 365.there are 148 instances 1st class and 217 instances is 2st class 1st class is the minor class because its error rate is 68/148 = 45.9% while the error rate of class 2 is 86/217 = 39.6% based on 10 fold cross validation in CART. The overall error rate is 42.75% six continuous attribute is class attribute that have value of 1 or 2.table 1 for attributes description.

Table 1
The meaning of Attributes

S. No.	*Attribute*	*Meaning*
1.	T4	Serum thyroxine
2.	FT4F	Free thyroxine fraction
3.	FT4	Free thyroxine
4.	THBR	Thyroid hormone binding ratio
5.	FT4I	Free thyroxime index
6.	T3	Serum triodothy roine
7.	FT3	Free triodothyroine
8.	FT3I	Free T3intex
9.	RAIU	Radioactive iodine uptake
10.	TSH	Serum thyrotropin
11.	TBG	Thyroxine-binding globulin
12.	TG	Thyroxine thyroglobalin

CART were used to generate decision trees for twelve random sample sets. Sample sets of size 180 were used. Remaining data were used for test

Table 2
The accuracy of decision tree By CART for sample sets

Sample Set #	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Con v	77.33%	67.72%	86.12%	66.38%	61.70%	66.96%	63.34%
Over samp 100%	69.50%	65.85%	66.29%	68.85%	84.14%	70.16%	70.23%
Over. samp 200%	67.80%	70.96%	NA	59.86%	68,63%	NA	69.85%

If the table 2 have the better results with over sampling in 4 out of 7.

5. CONCLUSION

Thyroid disorder at earlier stage different researchers have proposed different techniques to predict the thyroid disoreder and different kinds of accuracy level as per used techniques.so diagnosing thyroid disorder disease is a high interest to researchers of data miners, and decision trees have been a good data mining tools with respect to understandability and tranformability .but weakness of decision trees arises due to the fact that their branching criteria give higher priority for major classes.UCI thyroid disorder data set that is our interest for data mining is relatively small and has high error rate so that it may be vulnerable due to the property of decision trees

In order to overcome the problem of disdaining minority classes of the data set in decision tree generation algorithms, CART, showed very good results so that we may recommend oversampling for the data set to generate decision trees.future work is to see the effect of the method with smaller percentage of increase in minor class incrementally.

REFERENCES

[1] Deepika Koundle, Savita Gupta, Sukhwinder, "Computer-Aided Diagnosis of Tyroid Nodule: *A Review", omputer science & Engineering Survey (IJCSES),* August 2012.

[2] Nikita Singh, Alka Jindal , "A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images", *International Journal of Computer Applications* 0975–8887, July 2012.

[3] Nasrul Humaimi Mahmood and Akmal, "Segmentation and Area Measurement for Thyroid Ultrasound Image", *International Journal of Scientific & Engineering Research Volume 2*, December 2011.

[4] Preeti Aggarwal, Renu Vig, Sonali Bhadoria and C.G. Dethe, "Role of Segmentation in Medical Imaging: A Comparative Study", *International Journal of Computer Applications* 0975–8887, September 2011.

[5] Harris B, Othman S, Davies JA, Weppner GJ, Richards CJ, Newcombe RG *et al*. Association between postpartum thyroid dysfunction and thyroid antibodies and depression. *British Medical Journal* **305**, 152–156,1992.

[6] Eystraints G, "TND: A thyroid Nodule Detection System for analysis of Ultrasound Image and Videos", *Springer Science and Business Media*, LLC 2010.

[7] Mary C. Frates, Carol B. Benson, J.William Charboneau and Edmund S. "Management of Thyroid Nodules Detected at US: Society of Radiologists in US consensus", *conference statement management of thyroid nodules detected at US* **237(3)**.

[8] Prerana, Parveen Sehgal, Khushboo Taneja, "Predictive Data Mining for Diagnosis of thyroid Disease using Neural Network", *International Jurnal of Resaearch in Management, Scienc & Technology* 232-3264, **3(2)** April 2015.

[9] Ms. Priti Dhaygude, Mrs. S.M. Handore, "A review of thyroid disorder detection using medical images", *International Journal on Recent and Innovation Trends in Computing and Communication,* **2(12)**, December 2014.

[10] Prasanna Desikan, Kuo-Wei Hsu,JaideepSrivastava,"Data mining for healthcare management," *International Conference on data mining*, April 2011.

[11] Westberg, S Krogh, C Brink, and I R Vogelius, "A DICOM based radiotherapy plan database for research collaboration and reporting," *International Conference on the Use of Computers in Radiation Therapy* (ICCR 2013).

[12] Divdeep Singh Sukhpreet Kaur, "Scope of Data Mining in Medicine," *International Journal of Advanced Research in Computer Science and Software Engineering, Volume* 4, Issue 6, June 2014.

[13] Darsana. B., Dr. G. Jagajothi, "An Efficient DICOM Image Retrieval method based on features and neural network classification," *Journal of Theoretical and Applied Information Technology,* 31st October 2014.