Automatic Classification of Medical Data using Machine Learning

Tejaswini R. Patil[1] Shraddha A. Bhole[2] Yogita R. Bachhav[3] Snehal K. Bedase[4]

[1,2,3,4]NDMVP'S KBT COE

Abstract— It is evident that usage of machine learning methods in disease diagnosis has been increasing gradually. In this study, we will detect disease like Diabetes using ontology and genetic based machine learning approach. The first is an ontology-based classification that can directly incorporate human knowledge, while the second is genetic-based data mining algorithm that learns or extracts the domain knowledge from medical data in implicit form. The ontological modeling describes the relationship between classes and individuals. The Genetic-based data mining algorithm is a combination of genetic algorithm and machine learning tools for supervised learning which implement different classifiers. It is useful for analyzing the Diabetes disease with the help of symptoms. It is also better to suggest what kind of precautions the patient should take. In this way, these methodologies can be applied to help patients, students and physicians to decide the disease diabetes the patient has, what is the stage of disease and how it can be treated.

Key words: Ontology-based classification, Genetics-based classification, Data mining from medical data

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

It is evident that usage of machine learning methods in disease diagnosis has been increasing gradually. In this study, we will detect Diabetes disease using ontology and genetic based machine learning approach. The first is a ontology-based classification that can directly incorporate human knowledge, while the second is genetic-based data mining algorithm that learns or extracts the domain knowledge from medical data in implicit form. The ontological modelling describes the relationship between classes and individuals. The Genetic-based data mining algorithm is a combination of genetic algorithm and machine learning tools for supervised learning which implement different classifiers. It is useful for analyzing the disease type with the help of symptoms. It is also better to suggest what kind of precautions the patient should take. In this way, these methodologies can be applied to help patients, students and physicians to decide diabetes the patient has, what is the stage of disease and how it can be treated.

II. LITERATURE REVIEW

A. Machine Learning

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.

B. Existing System

The existing system is based on ontology and genetic based algorithm for epilepsy types. There are several DM techniques developed for diagnosing diseases. For example, Soni et al.[2] and Dangare and Apte[3] presented data mining techniques for heart disease diagnosis, and Ganesan et al.[4] presented the use of artificial neural networks for cancer diagnosis. In the authors designed surgical models of neurosurgery making use of ontology and described 106surgical cases. Through classification trees and clustering algorithms, they extracted surgical knowledge, facilitating the surgical decision-making process and surgical planning. In Lee et al. tried to overcome the limit of classical ontology dealing with uncertain knowledge and classified different diabetes syndromes using ontology-based inference rules.DM techniques are also developed for prognosing diseases [5]. Bayesian expert system for clinically detecting coronary artery disease is given. In [6] DM techniques are used for predicting heart attacks. In [7] artificial neural networks have been applied for prognosing end stage kidney disease. The work of Floyd [8] presents the application of DM techniques for prognosis of the pancreatic cancer. Moreover, some authors compared the performance of algorithms for the diagnosis or prognoses purposes. In [9], the discriminatory power of k-nearest neighbors, logistic regression, artificial neural networks, decision trees, and support vector machines on classifying pigmented skin lesions for diagnosis purpose is analyzed.

Limitations of existing system:

- The existing system determines only epilepsy types not all the diseases.
- It detects epilepsy types using ontology and genetic based method separately.

III. PROPOSED METHOD

A. System Architecture

The above fig. describes about system architecture of Automatic classification of medical data using machine learning. User has to register first, then he has to put the required information. There are the two methods for prediction of disease. First is Genetic based classification, in this method existing history of user is accepted. Second is Ontological based classification, here symptoms of diseases are accepted. From the analysis of this two method prediction of disease can be done.

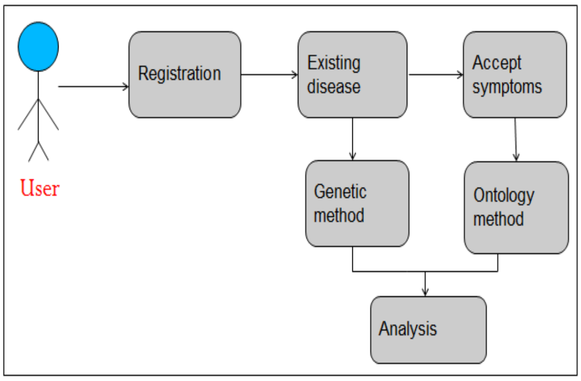

Fig. 1: Architecture Diagram

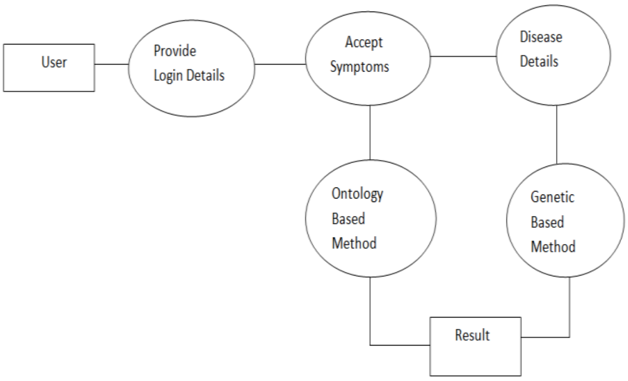

Fig. 2: Data flow diagram of the system

IV. CONCLUSION

This work aims at developing a fully automatic classifier for disease and their localization using symptoms and machine learning methods. Using this system diagnosis of disease like Diabetes or prediction of Diabetes disease is done without clinician concern.

V. FUTURE SCOPE

The proposed systems will be widely used in medical science.

REFERENCES

[1] Yohannes Kassahuna, Roberta Perroneb, Elena De Momib, Elmar Berghöferf, Laura Tassic, Maria Paola Canevinid, Roberto Spreaficoe, Giancarlo Ferrignob,Frank Kirchnera,faFachbereich: "Automatic classification of epilepsy types using ontology-based and genetics-based machine learning" 2014;

[2] Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis:an overview of heart disease prediction. Int J Comput Appl 2011;17(8):43–8.

[3] Dangare CS, Apte SS. Improved study of heart disease prediction system usingdata mining classification techniques. Int J Comput Appl 2012;47(10):44–8.

[4] Ganesan N, Venkatesh K, Rama MA, Palani AM. Application of neural networksin diagnosing cancer disease using demographic data. Int J Comput Appl2010;1(26):76–85.

[5] Chu C-M, Chien W-C, Lai C-H, Bludau H-B, Tschai H-J, LuPai S-MH, et al. ABayesian expert system for clinical detecting coronary artery disease. J MedSci 2009;4:187–94.

[6] Srinivas K, Rani BK, Govrdhan A, Karimnagar J. Applications of data miningtechniques in healthcare and prediction of heart attacks. Int J Comput Sci Eng2010;2(2):250–5.

[7] Di Noia T, Ostuni VC, Pesce F, Binetti G, Naso D, Schena FP, et al. An end stage kid-ney disease predictor based on an artificial neural networks ensemble. ExpertSyst Appl 2013;40(11):4438–45.

[8] Floyd S. Data mining techniques for prognosis in pancreatic cancer [Master'sthesis]. USA: Worcester Polytechnic Institute; 2007.

[9] Dreiseitl S, Ohno-machado L, Kittler H, Vinterbo S, Binder M. A comparison ofmachine learning methods for the diagnosis of pigmented skin lesions. J BiomedInform 2001;34:28–36.

[10] Siregar P, Toulouse P. Model-based diagnosis of brain disorders: a prototypeframework. Artif Intell Med 1995;7(4):315–42.

[11] Kohavi R. A study of cross-validation and bootstrap for accuracy estimationand model selection. In: Proceedings of the 14th international joint confer-ence on Artificial intelligence – vol. 2, IJCAI-95. San Francisco, CA, USA: MorganKaufmann Publishers Inc.; 1995. p. 1137–43.

[12] Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL plugin:an open development environment for semantic web applications. In: TheSemantic Web-ISWC 2004. Springer; 2004. p. 229–43.

[13] Noy NF, McGuinness DL. Ontology development 101: a guide to creating yourfirst ontology [Tech. rep.]. Stanford Knowledge Systems Laboratory (KSL-01-05)and Stanford Medical Informatics (SMI-2001-0880); 2001.

[14] Perone CS. Pyevolve: a python open-source framework for genetic algorithms.SIGEVOlution 2009;4(1):12–20.

[15] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKAdata mining software: an update. SIGKDD Explorations 2009;11(1):10–8.

[16] Curk T, Demar J, Xu Q, Leban G, Petrovic U, Bratko I, et al. Microarray data miningwith visual programming. Bioinformatics 2005;21:396–8.

[17] Ting K. Confusion matrix. In: Sammut C, Webb G, editors. Encyclopedia ofmachine learning. USA: Springer; 2010. p. 209.