ISSN:0975-9646

Prevention for Direct and Indirect Discrimination in Data Mining

Ms **Nikhita B Ugale**, Prof **S D Deshpande**

Computer Science and Engg. Department
Amravati University.
P.R.M.C.E.A.M. Badnera,
Amravati [MH] India.

Abstract : Data mining helps to extract useful and expected information among the huge amount of collective data present in database. Automated data collection with data mining collectively performs automated decisions. Discrimination can be direct or indirect. Direct discrimination use sensitive data for decision making. Indirect discrimination makes decisions on the basis of non-sensitive data. For more accuracy they express the relationship between discrimination prevention and privacy preservation in data mining. Along with security and privacy, proper discrimination performs vital role in considering legal as well as ethical point of view of data mining.

The main aim behind this paper is to develop new preprocessing discrimination prevention methodology which consist of different types of data transformation methods. With the help of that direct discrimination, indirect discrimination or both of them at the same time get prevented. For making the final decision there are two steps in which first step include identification of categories and makes groups of individuals whatever it may be, directly indirectly discriminated for making decision. In second step with the help of clustering, transformation of data in specific way such that removes all discrimination.

Keywords: **Direct discrimination , indirect discrimination, clustering etc.**

I. INTRODUCTION

With respect to the target concepts, applying some laws against discrimination, all of them are reactive, not proactive. Technology improve the accuracy factor by controlling discrimination. It required the large amounts of data. Those data are used to classify rules. Data mining tasks generates discriminatory models from biased data set as part of automated decision making. There may be direct discrimination or indirect discrimination, which is based on the sensitive discriminatory attributes related to group membership.

Basically, medical diagnosis is a decision making process, in which the physician induce the diagnosis of a new and unknown case from an available set of clinical data and from his/her clinical experience. To understand the progression of diseases and the efficacy of the associated therapies data mining techniques extract relationships and patterns holding in large amount of data.

Based on past data collected from the patients as well as the considerable symptoms of the patients system predict diseases. We focus on computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm which increases accuracy of such diagnosis by combining Neural Networks, Bayesian Classification and Differential Diagnosis all integrated into one single approach. With such advanced computing, doctors have always made use of technology to help them in various possible ways to take proper decision , from surgical imagery to X-ray photography. Unfortunately, technology has sometimes stayed behind when it came to diagnosis, a process that still requires a doctor's knowledge and experience to process the sheer number of variables involved, which involves medical history to climatic conditions, blood pressure, environment, and various other factors. Such accuracy yet not analyzed in computing technique. To overcome this problem, discrimination prevention in medical decision support systems helps to increase accuracy. Which will guides the doctors in taking correct decisions. Disease detection is a highly specialized and challenging job due to various factors, most of the times in case of diseases that show similar symptoms, or in case of rare diseases. This is analyze on the basis of their knowledge and experience, and it is later confirmed by performing related tests. Accuracy of the diseases is calculated upon clustering of symptoms, which involves detecting specific symptoms among all.

II. LITERATUER SURVEY

Shamsul I. Chowdhury [10] discusses in their paper issues related to the analysis and interpretation of medical data in 1994, thus allowing knowledge discovery in medical databases. He also explained that knowledge can also effectively gain from patient's database of observations and from interpretation of those observations. The important purpose behind that was to Study the feasibility of the approach exploring a large patient record system. The analysis was further carried out to test the hypothesis of a possible causation between hypertension and diabetes.

Hubert Kordylewski[11] , Daniel Graupe[11] explained the application of a large memory storage and retrieval (LAMSTAR) neural network to Medical diagnosis and medical information retrieval drawbacks in the year 2001. They also describes features of forgetting and of interpolation and extrapolation, which handle incomplete data sets. Applications of the network to three specific medical diagnosis problems are described: one is related to an emergency-room drug identification problem and anther two from nephrology.

Jenn-Lung Su, Guo-Zhen Wu [12] discuss the most widely used database concept in medical information system for processing large volumes of data in 2001. Symbolic and numeric data will define the need for new data analysis techniques and tools for knowledge discovery and decision making. Three popular algorithms for data mining which includes Bayesian Network (BN), C4.5 in Decision Tree (DT) , and Back Propagation Neural Network (BPN) were evaluated. The result shows that BN had a effective presentation in diagnosis ability.

Peter Kokol, Petra Povalej, Gregor Stiglic1, Dejan Dinevski [15] this paper presents the use of self organization to integrate different specialist's opinions generated by different intelligent classifier systems with having aim to improve classification accuracy in 2007. Fast and exact diagnosing of various diseases has proved that it performs vital role in health care processes. The main aim behind all is to mimic this real world situation in the manner to merge different opinions generated by different intelligent systems using the large amount of knowledge and self organizing abilities of cellular automata.

Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, Franco Turini [14] elaborates the Time- Annotated Sequence(TAS) concept in 2008. Time-annotated sequences(TAS), is a novel mining paradigm which solves this problem. TAS are sequential patterns where every transition between two events is annotated with a specific transition time that is found frequent in the data. The TAS mining paradigm is applied to clinical data regarding a set of patients in the follow-up of a liver transplantation. The main purpose of the data analysis is that of assessing the effectiveness.

Lishuang Li, Linmei Jing, Degen Huang [16] introduced a new way to extract Protein-Protein Interaction (PPI) information from biomedical literatures based on Support Vector Machine (SVM) and K Nearest Neighbors (KNN) in 2009. This setup is introduced to improve accuracy of SVM and KNN. To handle the unbalanced data distribution, a modified model of SVM-KNN classifier is proposed.

Demosthenes Akoumianakis, Giannis Milolidakis, Anargyros Akrivos, Zacharias [13] explain the concept of transformable boundary artifacts and their role in fostering knowledge-based work in cross-organization virtual communities of practice in 2010. Further on development through cycles of 'conception–elaboration– negotiation – reconstruction' provide number of facilities for clinical guideline development.

Rahul Isola, Rebeck Carvalho Amiya Kumar Tripathy [17] explained a system which uses Hopfield networks, LAMSTAR attempt which assist the doctors to perform differential diagnosis. In this system innovative utilization of the misdiagnosis factor for differentia diagnosis with that a possible method of implementation with the SOA technique in 2012. They introduced the important concept of Medi-Query in 2011. Instate of using medical data only for clinical and short term use, Medi-Query puts to use this vast knowledge of medical. so that diagnosis can be made on the basis of historical data.

Table no 1: Reference Papers Survey

Sr.	Year	Author	Advantages
1	1994	Shamsul I. Chowdhury Gustavsson R.	It explains a issue related to analysis & interpretation of medical data.
2	1999	Dr. Mirnmo Conforti	For early detection of cancer
3	2001	Hubert Kordylewski, Daniel Graupe	LAMSTAR is used for info retrival & also supplied interpolation & extrapolation of given data based on stored info.
4	2001	Jenn-Lung Su, Guo-Zhen Wu	It uses Bayesian n/w tech. for knowledge discovery. it gives specific accuracy in diagnosis of breast & tumor.
5	2007	Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, Franco Turini	It uses TAS Tech. It transplant solid organ without rejection.
6	2008	Peter Kokol, Petra Povalej, Gregor Stiglic1, Dejan Dinevski	To increase classification accuracy, it uses intelligent system.
7	2009	Lishuang Li, Linmei Jing, Degen Huang	They develop a new methods to extract Protein-Protein Interaction (PPI) information with the help of biomedical literatures based on Support Vector Machine (SVM) and K Nearest Neighbors (KNN).
8	2010	Demosthenes Akoumianakis, Giannis Milolidakis, Anargyros Akrivos, Zacharias	It gives us guideline management information system.
9	2011	Carvalho,Rahul Isola,Amiya Kumar Tripathy	It introduced a MediQuery, which help the medical fraternity in the long run by helping them in getting accurate diagnosis
10	2012	Carvalho, Rahul Isola, Amiya Kumar Tripathy	They introduce the new methods to differential diagnosis.

III. PROPOSED MODEL

In this model of medical application which prevents discrimination in the decision making system, clustering is used for accuracy and prevent confusion clearance which have similar symptoms. First it includes Training phase is a phase where we create a database by applying fuzzy rules on the various symptoms taken by doctors of patients. Our main aim behind this system is to identify earlier and accurate decisions about the diseases.

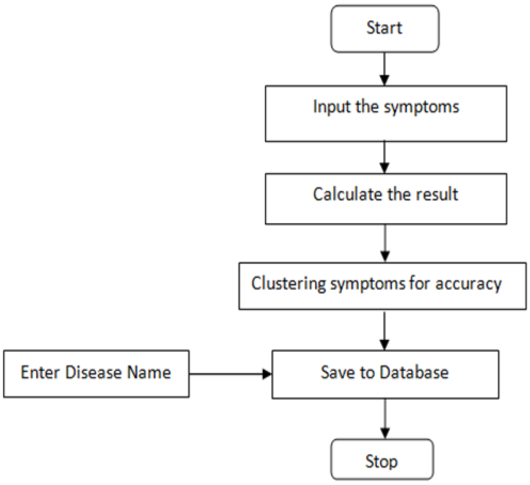

Fig- flow chart of proposed system

The proposed methodologies may contain 50 to 100 symptoms of various diseases the large data. In testing, we apply fuzzy rules on symptoms taken by the doctors, Pre-process it & out a one conclusion. Testing phase may include the following steps.

1. Collects symptoms from the patients.
2. Pre-process symptoms.
3. Extract features from database, apply fuzzy rules and clustering, which gives one proper conclusion.

IV. CONCLUSION:

Discrimination is a very important aspect when considering the legal and ethical issues of data mining along with privacy. For any personal reasons most of the people don't like discrimination in specific areas such as religion, nationality, age, and so on. But when it comes to the medical decisions there is no adjustment. The provided data should be accurate and proper so that it could give us the exact decision about medical issues. The main aim behind this paper was to develop a new preprocessing system that can prevent discrimination, which including different data transformation methods that can prevent. To prevent direct discrimination, indirect discrimination or both of them at the same time. To obtain this prevention techniques, the first step is to calculate discrimination exist and identify categories of those discrimination as well as the groups of individuals that have been directly and/or indirectly discriminated in the decision-making processes. The second step is to transform data in the proper way to remove all those discriminatory biases. At last, discrimination-free data models can be produced from the transformed data set without seriously damaging data quality. Knowledge is one of the most significant assets of any organization and especially in healthcare environment. clinical environment is at the top of information. However, getting specific information from received data is still a difficult challenge. Practical use of healthcare database systems, Decision making and management technologies like data mining can enormously contribute to increase accuracy in decision making in healthcare. Converting massive, complex and heterogeneous healthcare data into knowledge can help in controlling cost and maintaining high quality of patient care. A various models of data mining techniques have increasingly applied to tackle various difficulties and challenges of knowledge discovery in administrative and clinical facets of healthcare. In respect to clinical decisions, intelligent data mining tools can contribute effectively to enhance effectiveness of disease treatment and preventions as in the case of any critical diseases.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proc. 20th Int'l Conf. Very Large Data Bases,* pp. 487-499, 1994.

[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery, vol. 21, no. 2,* pp. 277-292, 2010.

[3] European Commission, "EU Directive 2004/113/EC on Anti-Discrimination," *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:EN:PDF,* 2004.

[4] European Commission, "EU Directive 2006/54/EC on Anti-Discrimination," *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:en:PDF,* 2006.

[5] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," *Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11),* pp. 47-54, 2011.

[6] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Rule Protection for Indirect Discrimination Prevention in Data Mining," *Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11),* pp. 211-222, 2011.

[7] F. Kamiran and T. Calders, "Classification without Discrimination," *Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09),* 2009.

[8] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," *Proc. 19th Machine Learning Conf. Belgium and The Netherlands,* 2010.

[9] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," *Proc. IEEE Int'l Conf. Data Mining (ICDM '10),* pp. 869-874, 2010.

[10] **S . I.** Chowdhury.: Statistical Expert Systems - A Special Application areafor Knowledge-based Computer Methodology. *Linkoping Studies in Scienceand Technology, Thesis No-104.,Department of Computer and Information Science, University of Linkoping, Sweden.*

[11] H. Kordylewski and D. Graupe, "Applications of the LAMSTAR neural network to medical and engineering diagnosis/fault detection," *in Proc7th Artificial Neural Networks in Eng. Conf., St. Louis, MO,* 1997.

[12] G.Z. Wu, "The application of data mining for medical database", *Master Thesis of Department of Biomedical Engineering, Chung Yuan University, Taiwan, Chung Li*, 2000.

[13] D. Akoumianakis, N. Vidakis, G. Vellis, D. Kotsalis, G. Milolidakis, A. Plemenos, A. Akrivos and D. Stefanakis, Transformable Boundary Artifacts for Knowledge-*based Work in Cross-organization Virtual Communities Spaces, Journal of Intelligent Decision Technologies Vol. 5 (1)*, 2011, in press.

[14] M. Berlingerio, F. B. F. Giannotti, and F. Turini, “Mining clinical data with a temporal dimension: A case study,” *in Proc. IEEE Int. Conf. Bioinf Biomed., Nov. 2–4, 2007,* pp. 429–436.

[15] Kokol P, Povalej, P., Lenič, M, Štiglic, G.: Building classifier cellular automata. 6th international conference on cellular automata for research and industry, *ACRI 2004, Amsterdam, The Netherlands, October 25-27, 2004. (Lecture notes in computer science, 3305). Berlin: Springer*, 2004, pp. 823-830.

[16] L. Li, L. Jing, and D. Huang, “Protein-protein interaction extraction *from biomedical literatures based on modified SVM-KNN,” in Nat. Lang. Process. Know. Engineer., 2009*, pp. 1–7.

[17] R. Carvalho, R. Isola, and A. Tripathy, “MediQuery—An automated decision support system,” *in Proc. 24th Int. Symp. Comput.-Based Med Syst.,Jun. 27–30, 2011,* pp. 1–6.