Identifying the Diagnosis of Diabetes Mellitus by using AK-Mode Algorithm

K. Elakkiya
M.Phil Scholar
Bharathidasan University

Abstract— Data mining aims at withdrawal of previously anonymous information from large databases. It can be observed as an automated solicitation of algorithms to discover hidden patterns and to extract information from data. Medicine is a new route in his undertaking is to prevent, diagnose and medicate diseases using data mining. Generally the data mining techniques, gathering and decision tree induction were used. Clustering is used to group patients according to the overall presence/absence of obliterations at the tested markers. Decision trees remained used to examine the resulting clustering and look for associations between deletion patterns, populations and the experimental of infertility. Decision support system to deliver Analytical Processing techniques are used to provide analysis of data. However, in order to integrate data mining results with data has to be demonstrated in a particular type of schema. The proposed work aims at the comparison of four algorithms called AK-mode algorithm, K-mode Algorithm, ROCK Algorithm, And MULIC Algorithm. Finally AK-mode Algorithm provides better results compared with the other algorithms.

Key words: Data Mining, Diabetic Approach, Clustering, AK-Mode Algorithm, Performance Evaluation

I. INTRODUCTION

At the present day world because of the lack of time number of the people avoids going through the large volume of database. The data warehousing is becoming more and more important in terms of considered to making the judgment through their competence to contribute assorted data from manifold information sources in a common storage space, for querying and analysis. The quality of services is important to deliver the healthcare Industry faces strong pressures and also reduce costs. Oftentimes, information produced is extreme, fragmented, imperfect, inaccurate, in the inaccurate position, or complicated to make good judgment [16]. A dangerous problem facing the industry is the lack of appropriate and timely information. These in sequence retrieval techniques allows to retrieve the large volume of database within compact point in time and in an simple format of the way the amount of citizens chooses these technique as a source of information retrieval techniques provides the Database Queries, Data Mining and Classification and Clustering techniques.

According to [7] & [8] systems have rapidly gained momentum in both the academic and research communities, mainly due to their fast and multi-dimensional investigation capabilities. In order to make easy this task propose the use of clustering as a data mining procedure to collection the dissimilar schemas resulting from the process of transforming the requirements.

II. PREVIOUS APPROACH

Diabetes is a defect in the body's ability to convert glucose (sugar) to energy. Glucose is the main source of fuel for our body. When food is digested it is changed into fats, protein, or carbohydrates. Foods that affect blood sugars are called carbohydrates. Carbohydrates, when digested, change to glucose. Individuals with diabetes should eat carbohydrates but must do so in moderation. Glucose is then transferred to the blood and is used by the cells for energy. In order for glucose to be transferred from the blood into the cells, the hormone - insulin is needed. Insulin is produced by the beta cells in the pancreas (the organ that produces insulin). In individuals with diabetes, this process is impaired.

By employing the analysis of big data will produce the predicted results for understanding the trends to improve the health care and life time expectancy, proper treatment at early stages at low cost.Due to the growing unstructured nature of diabetic data form health industry or all other sources, it is necessary to structure and emphasis its size into nominal value with possible solution.

III. PROPOSED APPROACH

A. AK-mode Algorithm

In this section propose AK-Mode which is an addition of the k-mode algorithm where we use the ontology to calculate the distinction distance. The Data Mining (DM) is "the examination of (often large) observational data sets to find unanticipated relationships and to recapitulate the data in novel ways that are both comprehensible and useful to the data owner". Many techniques and algorithms are used; in the following give some of them: gathering, cataloguing, calculation, etc.Clustering can be functional to various types of data: unbroken numerical variables, binary variables, categorical variables. In our case recommend its use to collection Requirement Schemas (RS).

Indeed, each RS is composed by a set of magnitudes, measures, fact and levels.

B. K-mode Algorithm

The k-modes approach adapts the standard k-means procedure for clustering categorical data by replacing the Euclidean detachment function with the simple corresponding dissimilarity measure, using modes to represent cluster centers and apprising modes with the most frequent resounding values in each of repetitions of the clustering process. These alterations guarantee that the clustering process meets to a local minimal result. Since the k-means gathering process is essentially not changed, the effectiveness of the clustering process is maintained.

C. ROCK Algorithm

It uses a combination of accidental sampling and divider clustering to handle large catalogs. In addition, its

hierarchical clustering algorithm symbolizes each cluster by a certain number of points that are engendered by selecting well dispersed points and then shrinking them toward the cluster centroid by a specified fraction.

D. MULIC Algorithm

Frequent item sets used to produce association rules are used to hypothesis a weighted hyper graph. Each frequent item set is a hyperactive edge in the subjective hyper graph and the weight of the hyper edge is subtracted as the average of the confidences for all conceivable association rules that can be engendered from the item set. Then, a hyper graph segregating algorithm from is used to partition the matters such that the sum of the weights of hyper edges that are cut due to the partitioning is minimized.

IV. IMPLEMENTATION METHODOLOGY

In this segment, based on our achievement discuss in details the stepladder concerned in the execution of the planned model using the conventional version of k-mode, this level is ignored. By applying association rule mining which is the way get better the effectiveness of this evaluation.

A. Data Set

The data set second-hand for the reason of this revision is Pima Indians Diabetes catalog of National Institute of Diabetes and Digestive and Kidney Diseases statistics sets are easy to get to in https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes this sites.

B. Data Cleaning

Data cleaning, as indicate, is in addition intimately associated to data mining, with the purpose of suggestive of possible inconsistency.

C. Clustering

Cluster exploration [6] divide data points into collection of points that are "close" to each one further. It starts with all data point being a cluster and frequently aggregates the most similar (least dissimilar) groups mutually in anticipation of attendance is just one big group. The numeral of groups can be chosen consequently.

D. Hierarchical Clustering of data

Hierarchal Clustering Explorer (HCE) apparatus is use for generating the hierarchical clusters of data. This tool takes key data file and allows the hierarchical clustering of agreed data based on disparate clustering parameters. At this point, consumer can select the parameters to present exact type of hierarchical clustering on the data.

V. LITERATURE REVIEW

The purpose of this literature review is to introduce and identify the limitations of automatic schema generation process by the other researchers. Our focus would be on the use of hierarchical clustering to automate the process of association rule mining schema generation.

RupaBagdi et al [2] developed a decision support system which combined the strengths of data mining process. This system would predict the future state and generate useful information for effective decision-making. They also compared the result of the ID3 and C4.5 decision tree algorithms. The system could discover hidden patterns in the data and it also enhanced real-time indicators and discovered bottlenecks and it improved information visualization.

N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik [23] Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. Data mining is a process which finds useful patterns from large amount of data. This technology has been successfully applied in Engineering and Technology, Science, Health Care Systems, Medical Diagnose Systems, Marketing and Finance to assist new discoveries and fortify markets. Some of the organizations have adapted this technology to progress their businesses and found outstanding results. In this paper we discussed a broad overview of some of the data mining techniques, their use in various emerging algorithms and applications. It provides an impression of the development of smart data analysis in medicine from a machine learning irrespective.

F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O [24] Modern electronic health records are designed to capture and render vast quantities of clinical data during the health care process. Technological advancements in the form of computer-based patient records software and personal computer hardware are making the collection of and access to health care data more manageable. However, few tools exist to evaluate and analyze this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. A common goal of the medical data mining is the detection of some kind of correlation, for example, between genetic features and phenotypes or between medical treatment and reaction of patients (Abidi & Goh, 1998; Li et al., 2005). The characteristics of clinical data, including issues of data availability and complex representation models, can make data mining applications challenging.

G. Parthiban, A. Rajesh, S.K.Srivatsa [29] The objective of our paper is to predict the chances of diabetic patient getting heart disease. In this study, we are applying Naïve Byes data mining classifier technique which produces an optimal prediction model using minimum training set. Data mining is the analysis step of the Knowledge Discovery in Databases process (KDD). Data mining involves use of techniques to find underlying structures and relationships in a large database. Diabetes is a set of related diseases in which body cannot regulate the amount of sugar specifically glucose (hyperglycemia) in the blood. The diagnosis of diseases is a vital role in medical field. Using diabetic's diagnosis, the proposed system predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease.

V.Karthikeyani, I.Parvin Begum, I.Shahina Begam K.Tajudin [33] Diabetes affects between 2% to 4% of the global population and its avoidance and effective treatment are undoubtedly crucial public issues in the 21st century. Although human decision making is often optimal, it is poor when there are huge amounts of data to be classified. Medical data mining has been a great potential for exploring

hidden patterns in the data sets of medical domain. Data mining algorithms can be trained in clinical data to predict the disease. Classification is the generally used technique in medical data mining. This paper presents results comparison of five supervised data mining algorithms using five performance criteria. The performance is evaluated by the five algorithms C4.5, SVM, k-NN, PNN, and BLR. Comparison of performance of data mining algorithms based on computing time, precision value, the data evaluated using 10 fold Cross Validation error rate, bootstrap validation and accuracy. A typical confusion matrix is furthermore displayed for quick check. The study describes algorithmic discussion of the dataset for the disease acquired from UCI and ICMR-INDIAB, on line repository of large datasets. Tanagra tool is used to achieve the best results. Tanagra is data mining matching set.

A. Sample Data Set

A knowledge discovery sample dataset is created to mine for two-year. The total dataset contains 768 instances. The following table shows the samples of the original dataset. It appearances the 9 attributes out of which diabetes probability is the class attribute. The other 7 attributes are used for decision making by C4.5 algorithm. The attributes used for diabetic prediction is ID, gender, Number of times conceived, plasma glucose, skin fold thickness, serum insulin, BMI, Diabetic type, Diabetic probability, Age, Blood pressure, other problems(like jaundice, TB, Sinuses, heart diseases etc).

ID	Sex	No. of. Times Pregnant	Plasma Glucose (mg/dL)	Diastolic B.P. (mm Hg)	Skin Fold thickness (mm)	2-Hr Serum Insulin (mu U/ml)	BMI (Kg/m^2)	Diabetes Pedigree Type	Diabetes Probability
1	M	-	160.50	59	30.75	142	29.35	2	High
2	M	-	98.30	68	35.75	66	27.75	1	Low
3	M	-	128.25	92	32.25	100	28.25	2	Medium
4	F	1	130.20	50	28.75	121	29.25	2	Medium
5	F	2	100.25	80	29.25	70	25.25	2	Low
6	M	-	110.35	86	36.25	73	29.75	2	Low
7	F	0	170.25	112	27.25	131	30.25	2	High

Table 1: Sample Data Set

VI. Result and Discussion

The report provides an analysis of more comprehensive and easier decision-making process through the allocation of doctors to under-represented geographic areas. It allows improving the quality of doctors in the areas of representation.

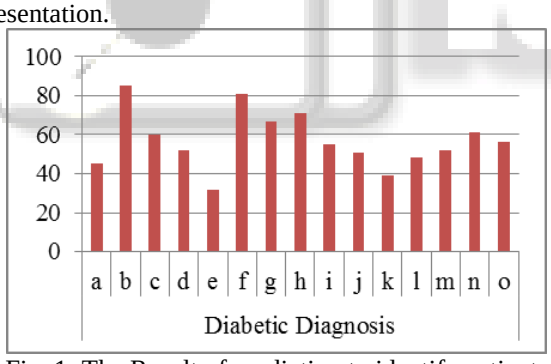

Fig. 1: The Result of prediction to identify patients

However, by combining, we can improve the current operations and to detect patterns more accurately in a time.

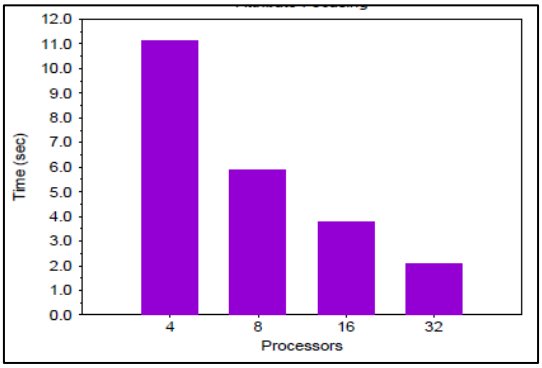

Fig. 2: Time for Consolidation query

With data mining and doctors can predict patients who may be diagnosed with diabetes. The result can enhance the previous processes and expose more subtle patterns, for example, by analyzing patient's demographics. Table demonstrations the result of prediction of a patient who was diagnosed as diabetic with high probability .The system was able to display this result in just 10 ms.

VII. Conclusion

This paper has obtainable a clinical DSS based on data mining with data mining to identify whether a patient can be analyzed with diabetes with likelihood high, low or medium. This is authoritative system because (1) it determines hidden patterns in the facts, (2) it improves real-time indicator and determine bottleneck and (3) it improves information conception. It is obvious from the result that the prototype system overcomes the physical plan design and execution prerequisite in the data warehousing environment. Further exertion can be done to enhance the system. For example, topographies can be added to allow doctors to query data cubes on business enquiries and automatically translate these questions to Multi-Dimensional eXpression (MDX) queries. The prototypical can also include composite data substances, spatial data and hypermedia data.

References

[1] V. Markl, F. Ramasak and R. Bayer, "Improving the performance by multidimensional hierarchical clustering," in Proc. of the 1999 Int'l Symposium on Database Engineering and Applications (IDEAS), 1999, p. 165.

[2] RupaBagdi, Prof. PramodPatil, "Diagnosis of Diabetes Using Data Mining Integration" in International Journal of Computer Science & Communication Networks,Vol 2(3), 314-322.

[3] R. Ben Messaoud, S. Rabaséda, O. Boussaid, and F.Bentayeb, "OpAC: A New Operator Based on a Data Mining Method", ixth International Baltic Conference on Databases and Information Systems (DB&IS 04), Riga, Latvia, 2004.
[4] Q. Chen, U. Dayal, and M. Hsu, "An Scalable Web Access Analysis Engine", In Proceeding of CASCON'97: Meeting of Minds, Toronto, Canada, 1997.
[5] V. Peralta, A. Marotta and R. Ruggia, "Towards the automation of data warehouse design," Technical Report TR-03-09, InCo, Universidad de la República, Montevideo, Uruguay, June 2003.
[6] Everitt B. (1980). Cluster Analysis (second edition). Halsted, New York.
[7] A. Omari, M. B. Lamine, and S. Conrad, "On Using Clustering And Classification During The Design Phase To Build Well-Structured Retail Websites", IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, 2008, pp. 51 59.
[8] A. Cuzzocrea, D. Sacca and P. Serafino, A hierarchy driven compression technique for advanced visualization of multidimensional data cubes, in Proc. of 8th Int'l Conf. on Data Warehousing and Knowledge Discovery (DaWak), (Springer Verlag 2006), pp. 106-119.
[9] Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering". *WIREs Data* Mining and Knowledge Discovery 1 (3): 231–240.doi:10.1002/widm.30
[10] D. Hand, H. Mannila and P. Smyth, "Principles of Data Mining", MIT Press, Cambridge, MA, 2001.
[11] VelidePhani Kumar, Lakshmi Velide, "A data mining approach for prediction and treatment of diabetes disease" in international journal of science inventions today Volume 3, Issue 1, January-February 2014.
[12] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[13] Panos, V., and Timos, S., A Survey on Logical Models for Diabetes Databases. ACM Sigmod Record, 28(4), 64-69, Dec. 1999.
[14] Hedger, S.R., The Data Gold Rush, Byte, 20(10), 83-88, 1995.
[15] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, 4(2), 9-17, March/April, 2002.
[16] Robert, S.C., Joseph, A.V. and David, B., Microsoft Data Warehousing: Building Distributed Decision Support Systems, London: Idea Group Publishing, 1999.
[17] Bill, G. F., Huigang, L. and Kem, P. K., Data Mining for the Health System Pharmacist. Hospital Pharmacy, 38(9), 845- 850, 2003.
[18] Usama F., Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDBM ˝97), Olympia, WA., 2-11, 1997.
[19] Raymond P.D., Knowledge Management as a Precursor Achieving Successful Information Systems in Complex Environments. Proceedings of SEARCC Conference 2004, 127-134, Kuala Lumpur, Malaysia.
[20] Usama, M. F., Data Mining and Knowledge Discovery: Making Sense Out of Data, IEEE Expert, 20-25, 1996, October.
[21] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, 4(2), 9-17, March/April, 2002.
[22] J. Han, and M. Kamber, 2006. Data Mining Concepts and Techniques, Elsevier Publishers.
[23] N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems" A Survey. International journal of computer applications, Vol. 2, No. 2, 2009.
[24] F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O (2009), "Challenges in Data Mining on Medical Databases", IGI Global, pp. 502-511.
[25] Raj Kumar, Dr. Rajesh Verma,‖ Classification Algorithms for Data Mining P: A Survey‖ IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.
[26] Breiman,L.,Friedman,J.,Olsen,R., Stone.C .,1984, ."Classification and Regression Trees", Chapman & Hall.
[27] J. Smola, B. Scholkopf, ‖A tutorial on support vector regression‖, Stat Comput 14 (2004) 199–222.
[28] Vidhya.K.A, G.Aghila‖ A Survey of Naïve Bayes Machine Learning approach in Text Document Classification‖ (IJCSIS) Vol. 7, No.2, 2010.
[29] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.
[30] Sarah Wild et al, Global prevalence of diabetes estimates for the year 2000 and projections for 2030, Diabetes Care, Vol. 27, No. 10, Oct. 2004, p. 25-60.
[31] Nitin Bhatia,Vandana,‖ Survey of Nearest Neighbor Techniques‖ (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.
[32] Charanjeet Kaur, ―Association Rule Mining using Apriori Algorithm: A Survey‖, IJARCET Volume 2, Issue 6, June 2013.
[33] V.Karthikeyani, I.Parvin Begum, I.Shahina Begam K.Tajudin, "Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction" , volume 60- No.12 December 2012
[34] K. R. Lakshmi and S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 ISSN 2229-5518

