Optimizing the Query Performance for Discovering the Diagnosis of Diabetics

K. Elakkiya
Research Scholar (M. Phil.)
Bharathidasan University

Abstract— Developing healthcare trade's move towards processing huge health records, and to admission entire for exploration and put into action will importantly increases the difficulties. Due to the growing unstructured nature of Big Data form health industry, it is essential to structure and emphasis its size into nominal value with possible solution. Health care management faces many challenges that make us to know the significance to develop the data analytics. Diabetic Mellitus (DM) one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. The extreme censure is to deal with large dataset with great amount of dimensionality, together in terms of the number of structures the data has, as well the number of rows of data that user is big business with. It can be observed as an automated solicitation of algorithms to discover hidden patterns and to extract information from data. Decision support system to deliver Analytical Processing techniques are used to provide analysis of data. 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. In this paper presents an integrative approach to conclude the diabetic disease from clinical big data. The clinical database is generally redundant, incomplete, dubious and unpredictable. The main objective of integrating is to experiment with different strategies of training data in order to increase the augury accuracy.

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

I. INTRODUCTION

Massive amount of information in healthcare allude to electronic health records, so vast and composite that they are not easy to maintain with outdated software system and/or hardware; nor will they be basically handled with outdated or mutual information managing implements and procedures. Huge records in healthcare is calamitous not only due to its size however more due to the variety of information forms and the haste at that it should be coped as it comprises speculative data and medicinal decision support systems.

Diabetes Mellitus commonly diabetes is a group of metabolic diseases in which a person has high blood sugar (blood glucose), either because the pancreas does not formed enough insulin, or because cells do not proceed to the insulin that is formed. Glucose builds up in the blood and element a condition that, if not controlled, can lead to serious health confusions and even death. The exposure of death for a person with diabetes is twice the risk of a person of smilingly age who does not have diabetes. Diabetes mellitus is a growing epidemic that induces 25.8 million people in the U.S. (8% of the population), and generally 7 million of them do not know they have the disease. Diabetes leads to significant medical complications including is heart disease, stroke, nephropathy, retinopathy, neuropathy and referral vascular disease. Early description of patients at risk of developing diabetes is a major healthcare need.

Applicable management of patients at risk with lifestyle changes and medications can decrease the risk of developing diabetes by 30% to 60%. Multiple risk factors have been described affecting a large portion of the population. For example, pre-diabetes (blood sugar levels above normal range but below the level of measure for diabetes) is present in almost 35% of the grown population and increases the absolute risk of diabetes 3 to 10 bend depending on the latency of additional associated risk factors such obesity, hypertension etc. Comprehensive medical management of this large segment of the population to prevent diabetes represents an oppressive burden to the healthcare system. In response to the acute to identify patients at high risk of diabetes early, various diabetes risk indices (risk scores) have been refined. Some of these indices (e.g. the Framingham score) gained acceptance in clinical practice and are used as direction in treatment: patients presenting high risk scores are consider more aggressively.

These scores only provide a quantification of the peril, they are not suggestive of the factors that may have caused the elevation of the risk. Furthermore, these scores utilize individual risk factors in an additive fashion without taking intercommunication among them into account. Diabetes is part of the metabolic syndrome, which is a constellation of diseases including hyperlipidemia (assent triglyceride and low HDL levels), hypertension (high blood pressure) and central over weight (with body mass index exceeding 30 kg/m2). These diseases interact with each other, with cardiac and vascular diseases and thus perceptive and modeling these synergy is important. 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 assorted information sources in a common storage space, for querying and analysis.

The quality of services is important to distribute 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 different kind of 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 transmuting the requirements.

II. RELATED WORK

The goal of data mining is to excerpt higher level information from an abundance of raw data. Association rules are a key tool used for this mission. An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are events. The rule states that with a certain probability, called the *confidence* of the rule, when X occurs in the given database so does Y. Association rules are connotations that comrade a set of potentially interacting conditions (e.g. high BMI and the existence of hypertension diagnosis) with elevated risk. The use of association rules is particularly beneficial, because in extension to quantifying the diabetes risk, they also readily provide the physician with a "justification", namely the associated set of circumstances.

This set of conditions can be used to guide treatment towards a more personalized and targeted defensive care or diabetes management. Let an item be a binary indicator signifying whether a patient acquire the corresponding risk factor. E.g. the item htn betoken whether the patient has been diagnosed with hypertension. Let X denotes the item matrix, which is a binary covariate matrix with rows representing patients and the columns representing items. An item set is a set of items: it intimate whether the corresponding risk factors are all present in the patient. If they are, the patient is said to be capped by the item set (or the item set applies to a patient). An association rule is of form $I \rightarrow J$, where I and J are twain item sets. The rule represents an connotation that if J is likely to apply to a lenient given that I apply.

The item set I is the antecedent and J is the consequent of the rule. The vitality and "significance" of the association is traditionally quantified through the abutment and confidence measures. In association rule mining, items do not tragedy accurate roles: there are no designated witch variables or outcome variables. In other words, any item can appear in the anterior of one rule and in the ensuing of another. Divining association rule mining represented the first departure from this paradigm by designating a specific item as an outcome. The consequent of the predictive association rules is always the nominated outcome item. Regressive association rules and quantitative association rules further bolster this paradigm allowing for a continuous upshot variable y to assist as the "consequent" of a rule.

III. PROPOSED APPROACH

In this paper the diabetic disease dataset is considered for execution, this data set is gathered from various Diabetic Care Centers at Erode. About diabetic patients' data were considered for this prediction and some of which is shown in Table 1. The inputs considered are Age, Fasting Plasma Glucose (FPG), Post Prandial Plasma Glucose (PPG) and the output is D-Diabetic Status.

Inputs			Output
Age	**FPG(MG/DL)**	**PPG(MG/DL)**	**D**
52	95	290	0.7
51	109	452	0.91
.	.	.	.
55	291	357	0.93

Table 1: Clinical Database of Diabetics Mellitus

The main ability of this algorithm is to automatically adjust the step length in order to speed up the convergence process. This algorithm is best in terms of accuracy and convergence speed.

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 probe of (often large) experimental 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 assured number of points that are engendered by selecting well dispersed points and then declining 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.

A clinical decision-support system is any computer program designed to help healthcare professionals to make clinical verdict. In a sense, any computer system that deals with clinical data or knowledge is invented to provide decision support. It is accordingly useful to consider three types of decision-support functions, feeding from

generalized to patient specific. Computer-based clinical decision support (CDS) can be defined as the use of the computer to bring relent knowledge to bear on the health care and wellbeing of a patient. AK-mode uses distinct prepossessing operations such as data cleaning, data transformation, data integration, hence, its output can serve as valuable data for data mining [3, 10]. AK-mode operations such as drilling, dicing, slicing, pivoting, Filtering enable users to cruise data flexibly, define relevant data sets, analyze data at different granularities and visualize results in different structures [11, 8, and 24]. Applying these operations can make data mining more exploratory.

E. Big Data and Diabetics

Big data is a term for data sets that are so large or complex that traditional data processing applications are deficient. Challenges include analysis, capture, data creation, search, sharing, storage, deportation, visualization, querying, renovate and information privacy. The term often refers simply to the use of divining analytics, user behavior analytics, or assured other advanced data analytics methods that extract value from data, and hardly to a particular size of data set. [2] Accuracy in big data may lead to more confident decision making, and recover decisions can result in greater operational efficiency, cost reduction and reduced risk.

F. Healthcare

Big data analytics has helped healthcare recover by providing personalized medicine and prescriptive analytics, clinical risk mediation and auguring analytics, waste and care variability rebate, automated external and internal reporting of patient data, regulated medical terms and patient registries and fragmented point solutions. The level of data achieve within healthcare systems is not paltry. With the added adoption of mhealth, eHealth and wearable technologies the volume of data will endure to increase. There is now an even greater need for such environments to pay greater assiduity to data and information quality. "Big data analytics explicit often means `dirty data' and the fraction of data erratum increases with data volume growth." Big data analytics extent is absurd and there is a violent need in health service for intelligent tools for accuracy and plausibility control and conduct of information missed.

IV. EXPERIMENTAL RESULTS

Fig. 1: Performance Evaluation of Proposed Algorithm with Existing Algorithms.

A knowledge discovery sample dataset is created to fund for two-year. The total dataset 768 exponents. The following table shows the samples of the initial dataset. It appearances the 9 attributes out of which diabetes probability is the class attribute. The another 7 attributes are used for verdict making by AK-mode 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 hazard, Age, Blood pressure, other problems(like jaundice, TB, Sinuses, heart diseases etc.).

Figure 1 shows that the AK-mode algorithm performance compared with the ROCK, MULIC, and K-Mode algorithms. Applying our methodology of mixing spacing AK-mode with survival analysis made a combinational sizable amount of (statistically significant) rules. Many of those rules square magnitude slight spinoff of every surrogate leading to the stupor of the clinical patterns underlying the rule set.

Figure 2 shows that the execution time by predicting diabetics is in milliseconds with their accuracy, sensitivity, and specificity. One remedy to the current defect, that constitutes the main focus of this work, is to summarize the rule set into a smaller set that's to brief. We first review the present rule set and info summarization ways, then adduce a generic framework that these ways fit into and finally, we tend to spread these methods in order that they will take monotonous outcome variable (the martingale residual in our case) under scruitiny.

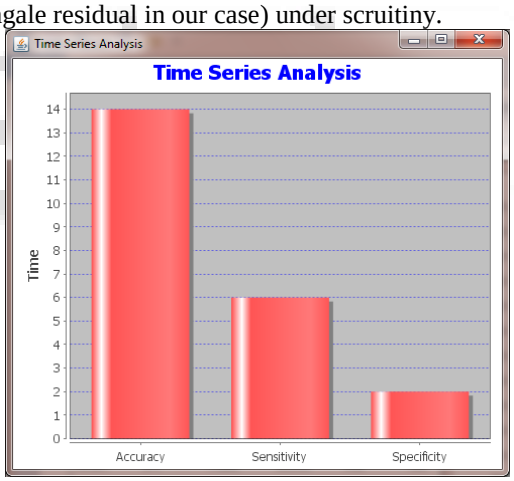

Fig. 2: Execution Time Analysis compared with as mentioned above Algorithms.

The key notability amidst the algorithms is that the definition of the loss benchmark. The loss paradigm is developed such that it absorb info respecting the expression of the rule further because the composed coverage of the rule. Unfortunately, with the viable exclusion of AK-mode, none of the strategies integrate associate outcome live like the risk of polygenic disorder.

V. CONCLUSION

This paper has obtainable a clinical DSS based on data mining with data mining to identify whether a enduring 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 axiomatically transcribe these questions to Multi-Dimensional eXpression (MDX) inquiry. 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, "AnScalable 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 UsingClustering And Classification During The Design Phase To Build Well-Structured Retail Websites", IADISEuropeanConference 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 Approachesand 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 IssueAugust 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 Surveyǁ, 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.

