The new frontier of Data mining techniques with proven intelligence-gathering technologies based on data enabled decisions in medicine

Prof.S.ANITHAA, VIT University, Chennai

Abstract

Healthcare is on the verge of a notable new period of discovery. With the advancement of the electronic health record, new opportunities for uncovering patterns will come to the forefront of medical knowledge. As the healthcare industry continues its drive to enhance quality of care, promote reliable services with reduction of cost, provided that these undiscovered patterns of care will become increasingly transparent for health care professionals. In recent times, the healthcare industry is beginning to apply the Evidence or Case of medical history which is purely based on proven intelligence-gathering technologies. Health care providers recognizes the value of data mining as a tool to analyze patient care and clinical outcomes but still lacking in the usage of information, and in fact it is far behind than other industries in creating integrated, longitudinal databases which can serve as repositories for data mining.

Introduction

In recent times there is rapid development in the area of information technology in biomedicine. To achieve rapid progress in mining medical data extensive studies have been made by researchers on various data mining tools. Researchers are increasingly looking forward the discovery of new techniques and innovative ideas in information technology that helps to overcome the rapid rise in health care costs faced by the community. Some of the application areas of data mining in highly visible fields like retail and marketing, e-business, fraud detection and in health care led to its application in Knowledge Discovery in Database**.** The objectives of this paper (i)is to enumerate current uses and highlight the importance of data mining in medicine and public health, and (ii)to improve the quality of data by using a novel pre processing technique (iii) To identify issues and Challenges in data mining as applied to the medical practice.(iv) To outline some recommendations for discovering knowledge in electronic databases through data mining.

Data Mining

Data Mining is a cross subject which covers machine learning, mathematical statistics, neutral network, database, pattern recognition, rough sets, fuzzy mathematics and relative technologies. The Data Mining in biomedical database can help us with disease diagnosis, treatment, research and decision making, by discovering the rules and mode of medical diagnosis. Data mining and its application to medicine and public health is a relatively young field of study. "Some authors refer to data mining as the process of acquiring information, whereas others refer to data mining as utilization of statistical techniques within the knowledge discovery process." The generally accepted definition of data mining today is the set of procedures and techniques for discovering and describing patterns and trends in data

Why Data Mining tools;

From the computerized health records there may be wealth of knowledge to be gained.. since there are very huge repositories, the amount of data stored in these databases makes it extremely difficult, and is not possible, for humans to analyze the data manually through it and to discover knowledge from them. Inorder .to overcome the complexity of present-day medical information the data mining tools are best-suited for this purpose.

Why Evidence-based medicine.

When medical professionals apply data mining techniques on Medical repositories they can discover new, useful and potentially life-saving knowledge, that otherwise would have remained inert in their databases. This knowledge can be stored and shared among health care professionals which leads to the prevention of hospital errors with respect to diagnosing disease..Since the knowledge learnt by data mining techniques is based on the experience of medical professionals it is referred as Evidence based or Case based learning

Early detection and/or prevention of diseases

By using data mining and visualization techniques, medical experts could find patterns and anomalies which is highly potential than traditional system and can be widely used in non- invasive diagnosis.

Invasive diagnosis.

Some diagnostic and laboratory procedures are invasive, costly, painful and intolerable to patients. An example of this is conducting a biopsy in women to detect cervical cancer.

Clinical Prediction rules (CPR) :

These rules are used by medical practitioners as formal guidelines in diagnosis, prognosis, and in treatment [1]. The rules simplify and expedite diagnosis and treatment for serious cases demanding immediate attention, and limit unnecessary diagnostic tests for low-probability cases. The rules provide quantitative predictive measures using factors from medical history, physical examination, and laboratory tests [2]. However, before the rules can be utilized in medical practice, they must be created, validated, and evaluated in clinical settings [3], [4]. By its nature, the creation of CPRs is time consuming and resource intensive. However, with the recent availability of electronic patient records and access to medical databases, the process of rule creation can be supported by machine learning methods providing automated or semi automated rule induction from data [5]–[7]. Such a data-driven approach reuses existing data sets collected for medical research and clinical records of diagnosed patients. The secondary use of medical data reduces the cost of data acquisition, provides access to rare medical cases, and allows for analysis of diversified populations. On the other hand, secondary analysis of data from heterogeneous sources presents several challenges [8], [9].

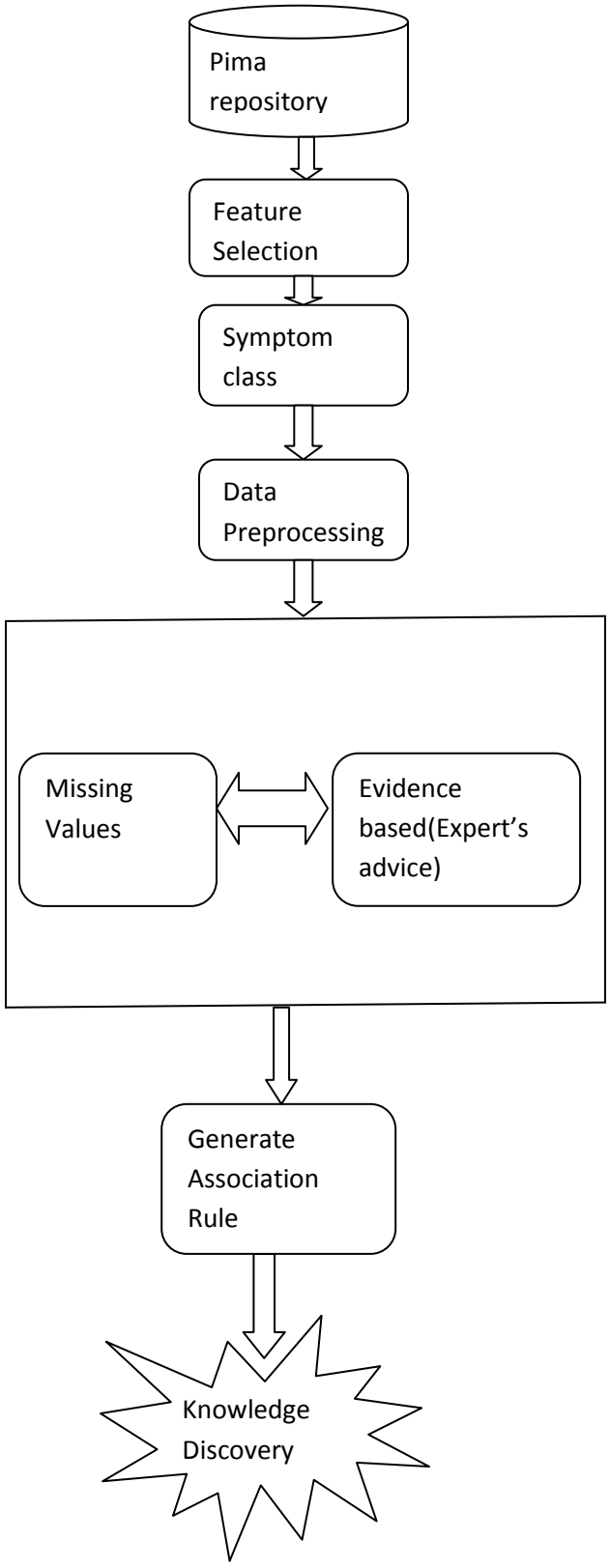
Model

In this model we used Pima Indian Diabetic data for analyzing the hidden knowledge in medical repositories As the first step the diabetic data from Pima repository is retrieved and the symptoms related to diabetic is fetched and the features with respect to the symptom were selected in order to create class called symptom class The symptom class is a cluster which consists of different symptoms of diabetic patient . From the symptom class each datum is preprocessed so that the missing values can be replaced by expert's knowledge . Later, a hybrid association rule mining algorithm is applied to the preprocessed data to generate the Rules with support and confidence.

Pima Dataset:

The, Pima Indian Diabetes dataset used was obtained from UCI machine learning repository .A study was conducted on 768 randomly selected female patients whose age was greater than 21. Characteristics of the patients like number of times of pregnancy and age in years ,plasma glucose concentration every 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps were recorded

Model



Data preprocessing

Data quality is a common issue that exists in many domains such as database systems, data mining, management information systems, In real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data we use Data preprocessing. In this process the parameters which has the value zero that is , zero valued instances are removed from the table and hence it was ended up with 625 instances out of 768.

TABLE I :PIMA INDIAN DIABETES DATA SET

s.no	Attributes	Description	Type
1	Pregnant	A record of the number of times the patient pregnant	Numeric
2	Plasma-Glucose	Plasma glucose concentration measured using a two-hour oral glucose tolerance test (mm Hg)	Numeric
3	Diastolic BP	Diastolic blood pressure	Numeric
4	Triceps SFT	Triceps skin fold thickness (mm)	Numeric
5	Serum–Insulin	Two-hour serum insulin (mu U/ml)	Numeric
6	*BMI*	Body mass index(weight Kg/height in $(mm)^2$	Numeric
7	*DPF*	Diabetes	Numeric

		pedigree function	
8	*Age*	Age of the patient (years)	Numeric
9	*Class*	Diabetes on set within five years	*Nominal*

Preprocessing techniques based on expert advice.

After the preprocessing only the 625 instances remain with 6 attribute age, pregnant, plasma glucose, diastolic BP, BMI, DPF and age. The following table summarizes the cut-off values along with the variable names:

TABLE II

S,no	Attributes	Range1	Range2	Range 3
1	**Pregnant**	: low (1,2)	medium (3,4,5)	high (> 6)
2	**Plasma – Glucose**	low (< 90)	medium (90–150)	high (> 150)
3	**Diastolic-BP**	normal (< 80)	normal-to-high (80–90)	high (> 90)
4	**BMI**:	low (< 25)	normal (25–30)	Severely bese (> 35)
5	**DPF**	low (< 0.4)	medium (0.4–0.8)	high (> 0.8)
6	**Age**	20–39(young)	), 40–59(medium)	60 plus (high)
7	**Class**	Yes(1)	No(0)	---

With respect to the above table all the 625 instances are converted into the categorical data which is shown in Table III.

TABLE III APPROXIMATION BASED ON EXPERT ADVISE

S.NO	**PREGNANT**	**PREG(CATEGORY)**	**PLASMAGLUCOSE**	**PLASMA GLUCOSE(CAT)**	**-**	**-**	**-**
1	**6**	**HIGH**	**148**	**MEDIUM**			
2	**1**	**LOW**	**85**	**LOW**	**-**	**-**	**-**
3	**8**	**HIGH**	**183**	**HIGH**			
4	**1**	**LOW**	**89**	**LOW**	**-**	**-**	**-**
5	**5**	**MEDIUM**	**116**	**MEDIUM**			
6	**3**	**MEDIUM**	**78**	**LOW**	**-**	**-**	**-**
7	**2**	**LOW**	**197**	**HIGH**			
8	**10**	**HIGH**	**168**	**HIGH**	**-**	**-**	**-**
-	**-**	**-**	**-**	**-**			
-	**-**	**-**	**-**	**-**	**-**	**-**	**-**

After conversion of continuous variable into the categorical variables Association rule mining algorithm- Apriori algorithm was used to find the hidden relationship between the variables. Apriori is an important association rule mining algorithm that is incorporated in many data mining software. Basically association rule works in two steps: (1) generating item sets that pass a minimum support threshold; and (2) generating rules that pass a minimum confidence threshold.

ALGORITHM APRIORI

1.Parameters given by users are UpperMinimumSupport, LowerMinimumSupport, Delta, Criterion, MinimumScore, and NumRules

2.Generate frequent itemsets that satisfy minimum support criterion

3.*Find frequent k-itemset, Sk, that satisfies the condition:*

LowerMinimumSupport ≤ *support(Sk)* ≤ *UpperMinimumSupport*

4.Generate rules that satisfy minimum confidence.

s.no	Rules	coverage	Confidence %
1	Plasma–Glucose= high and Age = 40–59 ⇒ yes	**60**	**84**
2	Plasma–Glucose= high and BMI =obes ⇒ yes	**56**	**78**
3	Plasma–Glucose =high and BMI = severely obese ⇒ yes	**66**	**82**
4	Pregnant = high and Plasma–Glucose =high ⇒ yes	**75**	**77**

5. Compute confidence
6. Sort set of association rules by Criterion

All the rules described here include all the combination and all risk factors that develop the diabetes within five year or not. Different association rules convey different regularities that trigger in the data set and generally predict the different things and so many association rule generated from even the data set is small. We keep such rules which are applicable reasonably large number of instances based on coverage and accuracy criteria. The *coverage* of an association rule is the number of instances for which it predicts correctly—this is often called its *support.* Its *accuracy* often called *confidence* is the number of instances that it predicts correctly, expressed as proportion of all instances to which it applies. The user has to specify the minimum coverage and accuracy values and look for only those rules whose values are at least of the specified minimum value. The Pima Indian diabetes data set would result in an enormous number of association rules, which would then have to be pruned down on the basis of their coverage and accuracy. Assuming that the minimum specified accuracy is 100%, only the first of these rules will make it into the final rule set. The algorithm, is implemented in WEKA, and will generate those top ten strongest rules for diabetes= yes which are shown in Table IV. Where as diabetes is equal = no, given in Table V. For example rule 1 the coverage is 74 and confidence is 100%.

TABLE-IV :RULE GENERATED BY THE ASSOCIATION MINING ALGORITHM FOR PATIENTS THOSE WHO BELONGS TO CLASS =YES

TABLE-V:RULE GENERATED BY THE ASSOCIATION MINING ALGORITHM FOR PATIENTS THOSE WHO BELONGS TO CLASS =NO

s.no	Rules	Coverage	Confidence %
1	Pregnant= low and diastolic Bp=normal and DPF=low and class=no⇒ age=20-39	74	100
2	Pregnant= low and pglucose=medium and DPF=low and class=no⇒ age=20-39	65	98

3	Pregnant = low and diastolic Bp=normal and age=20-39 ⇒ class=no	156	98
4	Pregnant = low and pglucose=medium and Diastolic Bp=normal and class=no⇒ age=20-39	113	97
5	pglucose=medium and BMI=low⇒ class=no	64	97
6	Pregnant = low and PF=low and class=no⇒ age=20-39	88	97
7	Pregnant = low and DPF= medium and class=no⇒ age=20-39	68	96
8	BMI=low and age20-39 ⇒ class=no	68	96
9	Pregnant = low and class=no⇒ age=20-39	181	96
10	Pregnant = low and pglucose=medium and class=no ⇒ age=20-39	123	95

CONCLUSIONS AND FUTURE WORK

In recent times, quality in healthcare is being shaped by evidence based medicine and the proper utilization of data. At its most basic level, quality is doing the right thing, at the right time, in the right way, for the right person. The challenges faced by clinicians every day is knowing , what the right thing is, when the right time is, and what is the right way. Knowing what data exists, understanding the data and making sense out of it can help clinicians rise to the challenge

In this case study, we used Evidence based preprocessing technique so as to improve the quality of data which is purely based on medical expert advice to Pima Indian diabetes data. Then we applied Hybrid Apriori association rule mining algorithm to generate the rules to find the hidden relationship among variables. By using this Model we can predict interesting patterns in pima data set which provides the information about the possibility of occurrences of diabetic in future. The future work will be focused on incorporation of outlier mining concepts in addition to evidence based preprocessing to enrich data quality.

REFERENCES

[1] M. H. Ebell, *Evidence-Based Diagnosis: A Handbook of Clinical Prediction Rules*. New York: Springer, 2001.
[2] A. Laupacis, N. Sekar, and I. G. Stiell, "Clinical prediction rules. A review and suggested modifications of methodological standards," *JAMA*, vol. 277, no. 6, pp. 488–494, Feb. 1997.
[3] I. G. Stiell, G. H. Greenberg, G. A. Wells, I. McDowell, A. A. Cwinn, N. A.

Smith, T. F. Cacciotti, andM. L. A. Sivilotti, "Prospective validation of a decision rule for the use of radiography in acute knee injures," *JAMA*, vol. 275, no. 8, pp. 611–615, Feb. 1996.
[4] T. McGinn, G. Guyatt, P. Wyer, D. Naylor, I. G. Stiell, and S. Richardson, "Users' guides to the medical literature XXII: How to use articles about clinical decision rules," *JAMA*, vol. 284, no. 1, pp. 79–84, 2000.
[5] W. Duch, R. Adamczak, K. Grabczewski, G. Zal, and Y. Hayashi, "Fuzzy and crisp logical rule extraction methods in application to medical data," in *Computational Intelligence and Applications*, P. Szczepaniak, Ed. Berlin, Germany: Springer-Verlag, 2000, pp. 593–616.
[6] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[7] A. Kusiak, J. A. Kern, K. H. Kernstine, and B. T. L. Tseng, "Autonomous decision-making: A data mining approach," *IEEE Trans. Inform. Technol. Biomed.*, vol. 4, no. 4, pp. 274–284, Dec. 2000.
[8] D. K. Owens and H. C. Sox, "Medical decision-making: Probabilistic medical reasoning," in *Medical Informatics: Computer Applications in Health Care and Biomedicine*, 2nd ed., E. H. Shortliffe and L.E. Perreault, Eds. New York: Springer, 2001, pp. 76–129.
[9] M. J. Druzdzel and F. J. Diez, "Combining knowledge from different sources in causal probabilistic models," *J. Mac. Learn. Res.*, vol. 4, pp. 295–316, 2003.

Author: Prof.S Anithaa VIT University Chennai campus To my credit I have published more than 12 papers in National and International journals and conferences. My area of interest are Data mining, Database management systems, Knowledge Engineering and applications, outlier mining and pattern recognition.