The 3rd International Conference on Mathematics and Statistics (ICoMS-3)
Institut Pertanian Bogor, Indonesia, 5-6 August 2008

IMPLEMENTATION OF CLASSIFICATION PREDICTIVE ASSOCIATION RULE (CPAR) ALGORITHM TO DIABETES DIAGNOSE

[1] Herwanto, [2] Imas S. Sitanggang

[1]Computer Science Faculty, University of Indonesia,
Kampus Baru UI, Depok – Indonesia

[2]Computer Science Department, Bogor Agricultural University
Jl. Meranti, Wing 20 Level V, Kampus IPB Darmaga, Bogor 16680 – Indonesia

e-mail : [1] herwanto_p@yahoo.com, [2] imasitanggang@yahoo.com

Abstract. *Hospital database yielded from hospital information system generally contains very much data with various attributes. Filtering and presenting relevant information in excessively database is difficult work. So needs certain techniques that screening of information can be done in efficient and effective, for example by applying data mining which will trace patterns from data for purpose of analysis. In this research studied how data mining can be applied to assist diabetes diagnose from data of medical laboratory.*
The real medical data set concerns patients with diabetes mellitus risk are included in diabetes data warehouse. Three steps are implemented for data mining process building. The first step is to deal with missing values. Next is the discretization step, where each variable is divided into a limited number of value groups. The next step is creating rule mining and classification.

There are 6.000 non-diabetes patients and 4.000 diabetes ones each with 12 variables: age, sex and results laboratory test. With Classification Predictive Association Rule (CPAR) algorithm, maximum predictive accuracy for diabetes is 69% and non-diabetes is 81%. The decision rules furthermore can used in application to predict diabetes or not. Data mining system building using CPAR algorithm is useful to diabetes diagnose.

Keywords : data mining, discretization, classification, Classification Predictive Association Rule algorithm

1. Introduction

Pattern of prevalence diabetes has shift. In the early of 1990 generally still confidence that diabetes only attack them which old age, and is " rich man disease". The present in reality diabetes has did not know class difference, diabetes can attack whosoever, either in " gedongan", slum, young and also old. Various genetic factors, environmental and way of living have a role in disease diabetes.
Various sighs can be met at diabetes. Classic sigh in the form of: polyuria, polyfagia and degradation of body weight. Other sigh can be in the form of: body weak, itchy and eye unclear. The diabetes Diagnose can be upheld with inspection of blood glucose at the time, and fasting blood glucose. In this research will be searched factors that relates with risk of diabetes from data result of inspection of laboratory tests.
The samples of research are based on database in Hospital Information System Rumah Sakit Pusat Pertamina (RSPP). Retrieval period of data from January 2005 until Decembers 2007. These data covers patients data which have diabetes risk number 2.430 patients. The selected patients are patients by minimizing once visit is estimated diabetes. The condition of medical note taken as sample refers to international classification of diseases tenth revision (ICD 10) where diabetes mellitus disease is given code E10 Insulin-dependent diabetes mellitus, E11 Non-insulin-dependent diabetes mellitus, E12 Malnutrition-related diabetes mellitus, E13 Other specified diabetes mellitus and E14 Unspecified diabetes mellitus.

There are 2.430 patients and have got 214.048 result of laboratory. To form training data and testing, taken 10 inspection types of laboratory tests which at most done that is Total Cholesterol (Chol), Triglyceride (Tg), Fasting Urine Glucose (Urn), Fast Urine Acetone (Actn), Fast Blood Glucose (Glun), HDL Cholesterol (Hdl), LDL Cholesterol (Ldl), Urine Glucose 2 hour PF(Upost), Urine Acetone 2 hour PF (Actpp).

Methodology of Classification model is based on three steps: a) handles incomplete data by means of extraction, b) changes continue data to discrete data and c) rule mining and classification. At first step, initial process of data diabetes is done to clean incomplete and extract data which will be applied to group disease diabetes or not. At second step is done transformation of continue data to discrete data. At third step, association rule mining algorithm is applied to produce rules which useful for detection of disease diabetes or not. According to Yin & Han (2003), an effective algorithm applying association rules for classification is the Classification based on Predictive Association Rules (CPAR). Forming of category values from data of laboratory tests in the early of process use the reference value from exist system in RSPP . To get model which better is done discretization process with more detail at attributes which becoming diagnosis determinant diabetes or not.

2. Data mining

Data mining is the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. Data mining can be told as process to filter or " mines" knowledge from a number of big data. There are three pillars of data mining represent three core areas of competency that are needed to be successful in data mining, that is; data mining techniques, data, and modeling of data. The techniques are general approaches to solving problems, and there are usually many ways to approach the technique. Each of these ways is a different algorithm. The main purpose of data mining generally to copes automatic process of decision with making model that having ability to do prediction or estimates a value. So factor that most important at model is accuracy.

The general steps of data mining consists of : 1) Data cleaning; to remove noise and inconsistent data. 2)Data Integration; where data from various sources is merged. 3) Data selection; where data relevant to the analysis task are retrieved from the database. 4) Data transformation; where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation. 5) Data mining; an essential process where intelligent methods are applied in order to extract data patterns. 6) Pattern evaluation; to identifies the truly interesting patterns representing knowledge based on some interestingness measures. 7) Knowledge presentation; where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2.1. Clustering

Clustering is the grouping of similar objects. Data object assumed by similar to stay in the same cluster. Purpose of clustering to lessen number of data by grouping similar objects into one groups. Object at clustering can show object physical or can also thing abstraction. Measurement which is made at the object called as feature. Pattern is a group of feature explaining object and in general representation as vector. Pattern clustering process typically involves the following steps :

1. Create a pattern representation, refers to number of classes, number of available patterns, and the number, type, and scale of the feature available to the clustering algorithm. The generation of pattern. representation is often dependent exclusively on the data available and the clustering algorithm being used However, it is sometimes possible and helpful to employ feature selection and/or feature extraction. Feature selection involves identifying the most effective features for discriminating patterns. Feature extraction is the process of transforming the available features to provide new features.
2. Define a pattern proximity measure appropriate to the data domain
3. Apply a clustering algorithm

K-means is one of clustering algorithm . It works best when the input data is primarily numeric. This algorithm divides a data set into a predetermined number of clusters. That number is the "k" in the phrase k means. A mean is just what a statistician call an average. In this case it refers to the average location of all of the members of a particular cluster. To form clusters , every record is mapped to a point in record space. this Space has as many dimensions as there are fields in the records. The value of each field is interpreted as a distance from the origin along the corresponding axis of the space. In order for this geometric interpretation to be useful, the fields must all be converted into numbers and the number must be normalized so that a change in one dimensions is comparable to a change in another

Records are assigned to clusters through an iterative process that starts with clusters centered at essentially random locations in the record spaces an moves the cluster centroid around until each one is actually at the center of some cluster of records.
In the first step, we select k data points to be the seeds. More or less arbitrarily. Every this seed is a embryonic cluster with only one element. In the second step., every record we give to cluster which centroid is nearest. After every point have been given to one cluster. The next step is to calculate centroids of the new cluster. This is simply a matter of averaging the positions of each point in the cluster along each dimension. In k-means clustering requires that the data values be numeric. Therefore, it is possible to calculate the average position just by taking the average of each field.

2.2. Classification

Classification is a form of data analysis than can be used to extract models describing important classes. Differed from clustering is made a pitch for class only from data, while at classification requires existence of data training in the form of class which have been definition before all. Data classification is a two –step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing databases tuples described by attribute. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. Since the class label of beach training sample is provided, this step is also known as supervised learning. In the second step, the model is used for classification. First, the prediction accuracy of the model (or classifier) is estimated, uses a test set of class labeled samples These sample are randomly selected and are independent of the training sample. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned model's class prediction for that sample. If the accuracy of the mode is acceptable, The model can be used to classify future data tuples or object for which the class label is not known.

2.3. Association rule

Association rule mining is a data mining technique and is a commonly used methodology for local-pattern discovery in unsupervised learning systems. Association rule mining is the discovery of association relationships or correlations among a set of items. They are often expressed in a rule form showing feature value conditions occurring frequently in a given dataset. An association rule in form X➔ Y is interpreted as database tuples that satisfy X are likely to satisfy Y. Tuples are rows in relational database while X and Y are composed of items and are called itemsets. X is called the antecedent part or body of the rule and Y the consequent part or head of the rule.
The most widely used framework for the evaluation of the association rules is the support confidence framework. The support of an association rule X➔ Y is the ratio of the tuples which contain the itemsets X and Y to the total number of tuples in the database. The confidence of an association rule X ➔ Y is the ratio of the tuples which contain the itemsets X and Y to the tuples which contain the itemset X Many efficient algorithms have been proposed for association rule mining, with the most common is the apriori algorithm. Problem from association rule mining can be divided in two step :

1. Find all frequent itemsets: By definition, each of these itemsets willoccur at least as frequnently as a pre-determined minimum support count.
2. Generate strong association rules from the frequent itemsets. By definition, these rules must satisfy minimum support and minimum confidence.

2.4. Algorithm apriori

The algorithm Apriori computes the frequent itemsets in database through several iterations. Each iteration has two steps; candidate generation and candidate counting and selection. In the first step of the first iteration, the generated sets of candidates item sets contains all 1-itemsets (item), these are all items in database. In the second step, this algorithm counts their support searching again through the whole database. Finally only i-itemsets (item) with s above required threshold will be selected as frequent. Thus, after the first iteration, all frequent i-itemset will be known.
In the second iteration, the algorithm Apriori reduce the set of candidate itemsets by pruning- a priori- those candidate itemsets that cannot be frequent. The pruning is based on the observation that if an itemset is frequent all its subsets could be frequent as well. Therefore, before entering the candidate counting step, the algorithm discards every candidate itemsets that has an infrequent subset.

2.5. Classification based association

A new field of data mining, classification using association rules, applies concepts used in association rule mining to classification problem. There are several methods which can be applied, for example association rule clustering system (ARCS) and associative classification. The method of ARCS mining association rule is based on clustering then applies rules to classification. The ARCS, mining association rule in the form of A_{quant1} A_{quant2} → A_{cat}, where A_{quant1} and A_{quant2} are data tests which their attribute has value spread, A_{cat} shows class label to category atribute given from training data.
Associative classification method mining rules in the form of condset → y, where condset is a group of items and y is class label. Rules satisfying with a minimum of certain support called as frequent. Rule had support s if s% from sample in set data containing condset and has class y. Rule satisfying with a minimum of confidence called as accurate. Rule had confidence c if c% from sample in set data containing condset to have class y. If some rule had the same condset, hence rule with highest confidence selected as possible rule (PR). This method applies algorithm association rule, like algorithm Apriori to result association rule, then chooses a group of rule having height quality and applies rules to predict data. But associative classification still less efficient because frequent yields more rules.
Method classification based other association is CPAR (Classification based on Predictive Association Rule). This algorithm takes idea from FOIL (First Order Inductive Leaner) in result rules and integrates it with associative classification.

2.6. Algorithm CPAR

The algorithm CPAR is started with read data in the form two dimensions array number which every the column is given attribute and the last attribute shows class. Input data furthermore is grouped to positive example P and negative example N as according to their classes. Weight of positive example | P| and negative example | N| for each attribute is summed up to form PN array, in the form of two dimension array contains list of all attributes, weight of positive example, and weight of negative example. The minimum weight threshold for P is calculated by multiplying the start weight of P by the Total Weight Threshold which was set to 0.05 during experimentation.
The process is done repeatedly until weight of positive example smaller than TWT. In each process is done copies P, N, A and PN to P', N', A ' and PN'. Calculates Gain and inserts rule to rule list. During experimentation minimum constant of gain is 07, and decay factor is 1/3. Gain is calculated based on formula :

$$Gain(p) = |P^*| \left[\log \frac{|P^*|}{|P^*| + |N^*|} - \log \frac{|P|}{|P| + |N|} \right]$$

where | P| are the positive examples and | N| are the negative examples satisfying the current rule r's body. | P*| are the positive example and | N*| are the negative examples satisfying the new rule's body. The best literal and the literal having similar gain are choosen to form the predictive association rule.
Laplace accuracy which good for knowing strength of prediction is calculated based on formula :

L.A = (nc+1) / (ntot+f)

where f is number of classes, ntot is the total number of examples satisfying the rule's body, among which nc example belong to c, which is the predicted class of the rule.

3. Result

3.1. Preprocessing phase

In the early step, data from relationship of patient master table, laboratory test and medical resume in the hospital database are transformed into working database, furthermore, are changed into form of category. Determining of category value for data of laboratory tests are based on reference as showed at Table 1. Determining of positive class of diabetes or negative diabetes determined by diagnose found in table of medical resume. The Class is specified as positive of diabetes if ICD code is E.10, E.11, E.12, E.13, or E14. The Category that are formed from the attributes are 29 categories, as showed at Table 2.

Table 1. The reference value of laboratory tests

Examination Code	Explanation	Unit	Normal Value

Actn	Fast Urine Acetone		<=0
Actpp	Urine Acetone 2 hour PF		<=0
Chol	Total Cholesterol	Mg/dL	< 200
Glun	Fast Blood Glucose	Mg/dL	70 ~ 110
Gpost	Blood Glucose 2 hour PF	Mg/dL	100 ~ 140
Hdl	HDL Cholesterol	Mg/dL	40 ~ 60
Ldl	LDL Cholesterol	Mg/dL	< 130
Tg	Triglyceride	Mg/dL	50 ~ 150
Upost	Urine Glucose 2 hour PF		<=0
Urn	Fasting Urine Glucose		<=0

Table 2. Data Category

Attribute	Continue Value	Category
Age	Age < 20	1
Age	20<= Age <= 40	2
Age	Age > 40	3
Sex	Sex = Male	4
Sex	Sex = Female	5
Glun	Glun < 70	6
Glun	70 <= Glun < 110	7
Glun	Glun >= 110	8
Gpost	Gpost < 100	9
Gpost	110<= Gpost <140	10
Gpost	Gpost >= 140	11
Upost	Upost <= 0	12
Upost	Upost > 0	13
Actn	Actn <= 0	14
Actn	Actn > 0	15

Attribute	Continue Value	Category
Urn	Urn <= 0	16
Urn	Urn > 0	17
Actpp	Actpp <= 0	18
Actpp	Actpp > 0	19
Ldl	Ldl< 130	20
Ldl	Ldl>=130	21
Hdl	Hdl < 40	22
Hdl	40<= Hdl< 60	23
Hdl	Hdl>=60	24
Chol	Chol<200	25
Chol	Chol>=200	26
Tg	Tg<50	27
Tg	50<=Tg< 150	28
Tg	Tg>=150	29

3.2. Processing phase

The processing of data model uses algorithm CPAR, result of process is compared to algorithm C50. Training is done by taking 8.000 samples, with proportion of positive diabetes and negative diabetes are 1:1. Data samples consist of information about input data, there are age, sex, result of laboratory test, output data in the form of diagnose disease. 8.000 data samples having complete medical note is collected. Data sample characteristic showed at Tabel 3, and Tabel 4. Result of data process mining of 8.000 data samples training with Gain similarity ratio 50% is showed at Tabel. 5.

Tabel 3. General characteristic of data

Data	Sex	Number of Rows	Percentage	Age		
				Mean	±	SD
Positive Diabetes	Male	2.306	28.63	58	±	14
	Female	1.694	21.18	56	±	16
Negative Diabetes	Male	2.506	31.33	54	±	23
	Female	1.494	37.43	59	±	14

Tabel 4. Average variabel of result laboratory tests

	Positive Diabetes			Negative Diabetes		
Inspection types	Mean	±	SD	Mean	±	SD
Fast Blood Glucose (mg/dl)	157.72	±	77.83	107.12	±	58.83
Blood Glucose 2 hour PF (mg/dl)	213.87	±	101.61	145.88	±	90.35
Urine Glucose 2 hour PF	1.50	±	1.40	0.42	±	0.95
Fast Urine Acetone	0.05	±	0.23	0.03	±	0.25

Fasting Urine Glucose	0.57	±	1.10	0.17	±	0.62
Urine Acetone 2 hour PF	0.03	±	0.19	0.02	±	0.19
LDL Cholesterol (mg/dl)	124.75	±	39.31	128.03	±	37.64
HDL Cholesterol (mg/dl)	51.97	±	18.55	56.23	±	26.61
Total Cholesterol (mg/dl)	205.30	±	49.51	199.80	±	50.06
Triglyceride (mg/dl)	172.40	±	113.80	147.90	±	91.80

Tabel 5. Rules with Gain similarity ratio 50%

No	Rules	L.A
1	IF Glun>=110 Then positive Diabetes	0.69
2	IF Gpost >= 140 Then Positive Diabetes	0.68
3	IF Upost > 0 Then Positive Diabetes	0.68
4	IF 70 <= Glun < 110 Then Negative Diabetes	0.72
5	IF Upost <= 0 Then Negative Diabetes	0.70
6	IF 100<=Gpost <140 Then Negative Diabetes	0.61
7	IF Gpost < 100 Then Negative Diabetes	0.81

Data in Tabel 5. shows that rule IF Glun >= 110 Then Positif Diabetes has Laplace Accuracy (LA) highest at class of positive diabetes that is 69%. This means that patient with result of inspection Glun bigger or equal to 110 having opportunity is hit diabetes equal to 69%. Rule IF GPOST < 100 Then Negative Diabetes has highest LA at class of negative diabetes that is 81%. Based on data at Tabel 5 seen also that inspection Gpost, Upost, Glun becomes main determinant to determine is positive diabetes or negative diabetes.

By using algorithm C50 is produced 13 rules, contain 6 rules of positive diabetes and 7 rules of negative diabetes. Rules for positive diabetes with highest confident is :

Positive IF GLUN>=110 THEN of Diabetes (0.47, 069)

IF GPOST >= 140 Positive Then of Diabetes (0.47,0.674)

Rules for negative diabetes with highest confident is :

If GLUN 70 <= Glun < 110 and Gpost < 100 and Urn <= 0 then Negative Diabetes (01, 089)

If Gpost < 100 and Urn <= 0 and 50<=Tg< 150 then Negative Diabetes (01, 088)

If GLUN 70 <= Glun < 110 and Ldl>=130 then Negative Diabetes (01, 087)

3.3. Improvement process

Improvement process is done with changing category Glun and Gpost becoming main determinant of positive diabetes or negative diabetes. The category that are formed from improvement process are 31 categories, as showed at Table 6. The rules that are resulted after improvement process with Gain similarity ratio 50% showed at Tabel 7.

Table 6. Data Category of improvement process

Atribut	Nilai Kontinyu	Katagori
Umur	Umur < 20	1
Umur	20<= Umur<= 40	2
Umur	Umur > 40	3
Sex	Sex = Laki-Laki	4
Sex	Sex = Perempuan	5
Glun	Glun < 70	6
Glun	70 <= Glun < 110	7
Glun	110 <= Glun < 140	8
Glun	Glun >= 140	9
Gpost	Gpost < 100	10
Gpost	100<= Gpost <140	11
Gpost	140 <= Gpost < 200	12

Atribut	Nilai Kontinyu	Katagori
Actn	Actn <= 0	16
Actn	Actn > 0	17
Urn	Urn <= 0	18
Urn	Urn > 0	19
Actpp	Actpp <= 0	20
Actpp	Actpp > 0	21
Ldl	Ldl< 130	22
Ldl	Ldl>=130	23
Hdl	Hdl < 40	24
Hdl	40<= Hdl< 60	25
Hdl	Hdl>=60	26
Chol	Chol<200	27

Gpost	Gpost >= 200	13		Chol	Chol>=200	28
Upost	Upost <= 0	14		Tg	Tg<50	29
Upost	Upost > 0	15		Tg	50<=Tg< 150	30
				Tg	Tg>=150	31

Tabel 7. Rules of improvement process with Gain similarity ratio 50%

No	Aturan	L.A
1	IF Glun >= 140 AND IF Gpost >= 200 Then Positif Diabetes	0.7 0.7
2	IF Upost > 0 Then Positif Diabetes	0.7
3	IF Upost <= 0 AND IF 70<=Glun<110 Then Negatif Diabetes	0.72 0.70
4	IF 100<=Gpost<140 Then Negatif Diabetes	0.61
5	IF Gpost < 100 Then Negatif Diabetes	0.81

4. Conclusions

The conclusion from this research are:

1. Inspection of blood glucose 2 hour pf (Gpost), urine glucose 2 hour pf (Upost), fast blood glucose (Glun) becomes main determinant to determine is patient positive diabetes or negative diabetes.
2. Algorithm CPAR only choose attribute value having best Gain value, so there are possibility that attribute having strength of high prediction doesn't emerge in rule.
3. Algorithm CPAR receives input in the form of category, causing determinate process of continue data to category data hardly having an effect on to result of prediction.

5. References

Berry MJ, Linoff GS, 2000, *Mastering Data mining : The Art and science of Customer Relationship Management*, New York: John Wiley & Sons, Inc.

Berson A., Smith S, Thearling K, 2001, *Building Data mining Application for CRM*, mcGraw-Hill.

Coenen F, 2004, *The LUCS-KDD Implementations of CPAR (Classification Based on Predictive Association Rules)*, Department of Computer Science The University of Liverpool..

Corey M, Abbey M, Abramson I, Taub B, 2001, *Oracle 8i : Data Warehousing*, Osborne/McGraw-Hill.

Jiawei H,M Kamber, 2001, *Data Mining Concepts and Techniques*, The Morgan Kaufmann Publishers

Joseph L. Breault, *Data Mining Diabetic Database: Are Rough Sets a Useful Addition?*, Departmnt of Health Systems Management, Tulane University.

Shital S, Adrew K, B , Dixon, *Data Mining in Predicting Survival of Kidney Dialysis Patients Invariant object approach*, Intelligent Systems Laboratory, 2003.

Soegondo S, Soewondo P, Subekti I, 2002, *Penatalaksanaan Diabetes Melitus Terpadu*, Jakarta: Balai Penerbit FK-UI.

Yin X, Han J, 2003, CPAR: *Classification based on Predictive Association Rules*, University of Illinois at Urbana-Champaign.