COMPARISON A PERFORMANCE OF DATA MINING ALGORITHMS IN PREDICTION OF DIABETES DISEASE

SHAIK.CHANDINI
B.TECH –CSE
SVCN- Nellore, Andhra Pradesh.
Email Id: chandinishaik2510@gmail.com

T.GAYATHRI NAGALAKSHMI
B.TECH –CSE
SVCN- Nellore, Andhra Pradesh.
Email Id: 313Gayathri@gmail.com

K.MAHENDRA
Assistant Professor in the dept. of CSE, SVCN,
Nellore, Andhra Pradesh,
EmailId: mahiknlr@gmail.com

ABSTRACT:

Detection of knowledge patterns in clinical data through data mining. Data mining algorithms can be trained from past examples in clinical data and model the frequent times non-linear relationships between the independent and dependent variables. The consequential model represents formal knowledge, which can often make available a good analytic judgment. Classification is the generally used technique in medical data mining. This paper presents results comparison of ten supervised data mining algorithms using five performance criteria. We evaluate the performance for C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA and PLS-LDA then Comparison a performance of data mining algorithms based on computing time, precision value , the data evaluated using 10 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative, 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, on line repository of large datasets. The Best results are achieved by using Tanagra tool. Tanagra is data mining matching set. The accuracy is calculate based on addition of true positive and true negative followed by the division of all possibilities.

KEYWORDS:

C4.5, SVM, K- NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA, PLS-LDA, Classification based on CT, Precision value, CV error rate, BV error rate and Accuracy.

INTRODUCTION

Basically declared, data mining refers to extracting or "mining" knowledge from large amounts of data or databases [1]. The development of finding useful patterns or importance in raw data has been called KDD (knowledge discovery in databases) [2]. Bulky number of data mining algorithms has been developed in modern days for mining of knowledge in databases. Of these many are supervised learning algorithms. These algorithms are generally used for categorization tasks. The importance in systems for independent decisions in medical and manufacturing applications is increasing, as data becomes available. In the previous century, an exponential inhancement has been seen in the accuracy and sensitivity of diagnostic tests, from observe outside symptom and use refined laboratory tests and difficult imaging methods increasingly that allow detailed non-invasive inner examinations. This improved accuracy has predictably resulted in an exponential increase in the patient data available to the physician. The process of finding confirmation to decide a probable reason of patient's key

symptoms from all other possible reason of the symptom are known as establishing a medical diagnosis.

The utilize of computer tools in medical decision support is now well-known and pervasive across a wide range of medical area such as diabetes, cancer etc.

Data mining is a remarkable opportunity to support physician deal with this large amount of data. Its methods can help physicians in various ways such as interpret multifaceted diagnostic tests, combining information from several sources (sample movies, images, clinical data, proteomics and scientific knowledge), given that support for differential diagnosis and providing patient-specific prediction. The respite of the paper is organized as follows: It first gives details of classification on different methods. Then medical data mining is described. In section IV, some instances of the prediction and diagnosis problems in medicine in case of diabetes diseases are considered. The article ends by concluding with a summary of investigated methods and future research.

LITERAURE REVIEW

There are diverse kinds of studies for DM techniques in medical databases.

[i] J.W.Smith et al dealing with this data base uses an adaptive learning routine that generates and executes digital analogs of perceptron-like devices, called ADAP. They used 576 training instances and obtained a classification of 76% on the remaining 192 instances. Classification is the most widely used technique in medical data mining.

[ii] J.L.Brute et al experimental results showed based on a measure of glycemic control related to outcomes. Used the classification tree approach in Classification and Regression Trees (CART) with a binary target variable of HgbA1c >9.5 and 10 predictors.

[iii] M.S.Miroslay et al using data mining technologies for prediction using classification accuracy adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables.

[iv] Milan kumari et al using four classification algorithms in prediction cardiovascular disease only concentrating accuracy.

[v] B.N.Prathiba et al using data mining SVM classifier in combined with transform domain mainly concentrating on domain accuracy.

DIFFERENT TASK OF DATA MINING

There are various numbers of data mining methods. One approach to categorize different data mining methods is based on their function ability as below.

1) **Regression** is a statistical methodology that is often used for numeric prediction.
2) **Association** returns affinities of a set of records.
3) **Sequential pattern** function searches for frequent subsequences in a sequence dataset, where a sequence records an ordering of events.
4) **Summarization** is to make compact description for a subset of data.
5) **Classification** maps a data item into one of the predefined classes.
6) **Clustering** identifies a finite set of categories to describe the data.
7) **Dependency modeling** describes significant dependencies between variables.
8) **Change and deviation detection** is to discover the most significant changes in the data by using previously measured values.

Classification algorithms require that the classes be defined based on data attribute values. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes. Data classification is a two-step process.

Step 1: A classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. Each tuple is assumed to belong to a predefined class called the class label attribute. Because the class label of each training tuple is provided, this step is also known as **supervised learning.** The first step can also be viewed as the learning of a mapping or function, y = f (X), that can predict the associated class label y of a given tuple X. Typically, this mapping is represented in the form of classification rules, decision trees, or mathematical formulae.

Step 2: The model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to over fit the data.

MACHINE LEARNING APPROACHES

Machine learning algorithms can be classified as supervised learning or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. Learning algorithms aim to predict output values of new examples based on their input values. In unsupervised learning, training examples contain only the input patterns and no explicit target output is associated with each input[13]. The unsupervised learning algorithms need to use the input values to discover meaningful associations or patterns.

In supervised machine learning algorithms (*C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA, PLS-LDA*).

EVALUATION METHODLOGIES

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task (Kohavi, 1995) . Choosing a good evaluation methodology is very important for machine learning systems development. There are several popular methods used for such evaluation, including holdout sampling, cross validation, leave-one out, and bootstrap sampling (Stone, 1974; Efron and Tibshirani, 1993). In the holdout method, data are divided into a training set and a testing set. Usually 2/3 of the data are assigned to the training set and 1/3 to the testing set. After the system is trained by the training set data, the system predicts the output value of each instance in the testing set. These values are then compared with the real output values to determine accuracy. In cross validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The accuracy of the system is the average accuracy over the 10 iterations. In the bootstrap method, *n* independent random samples are taken from the original data set of size *n*. Because the samples are taken with replacement, the number of unique instances will be less than *n*.

These samples are then used as the training set for the learning system, and the remaining data that have not been sampled are used to test the system (Efron and Tibshirani, 1993).

RESEARCH FINDINGS

Data mining in the diabetes disease Prediction

Ten different supervised classification algorithms i.e. C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA, PLS-LDA have been used analyze dataset in. Tanagra tool is powerful system that contains clustering, supervised learning, Meta supervised learning, feature selection, and data visualization supervised learning assessment, statistics, and feature selection and construction algorithms.

DATA SOURCE

To evaluate these data mining classification Pima Indian Diabetes Dataset was used. The dataset has 9 attributes and 768 instances.

Table 1. Attributes of diabetes dataset

No	Name	Description
1	Pregnancy	Number of times pregnant
2	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Pres	Diastolic blood pressure (mm Hg)
4	Skin	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	Mass	Body mass index (weight in kg/(height in m)^2)
7	Pedi	Diabetes pedigree function
8	Age	Age (years)
9	Class	Class variable (0 or 1)

Attributes are exacting, all patients now are females at least 21 years old of Pima Indian heritage. If the 2 hour post load Plasma glucose was as a minimum 200 mg/dl.

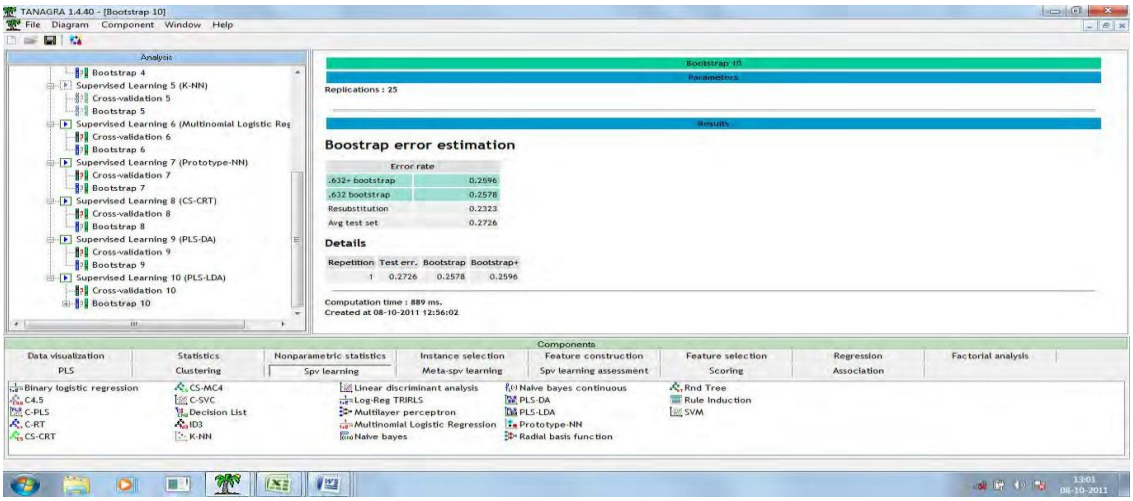


Figure 1: Screen shot for Bootstrap Validation Error rate Performance (10 Algorithms).

PERFORMANCE STUDY OF ALGORITHMS

The table 2 consists of values of different classification. According to these values the lowest computing time (<550ms) can be determined.

Table 2: Comparison of supervised Algorithms based on performance

							Acc	Spec	Sen	CV	P	N	
s.no	Algo	CTime	TP	FN	FP	TN	%	%	%	Erate	(Prec)	(Prec)	BVErate
1	C4.5	550ms	31	23	19	77	72	80%	57%	0.28	0.38	0.23	0.3196
2	SVM	546ms	24	30	14	82	70.667	85%	44%	0.29	0.368	0.268	0.2929
3	K-NN	640ms	20	34	18	78	65.333	81%	37%	0.34	0.474	0.304	0.3532
4	PNN	546ms	42	12	39	57	66	59%	78%	0.34	0.482	0.174	0.3406
5	BLR	515ms	32	22	19	77	72.667	80%	59%	0.273	0.373	0.222	0.2754
6	MLR	530ms	32	22	19	77	72.667	80%	59%	0.273	0.373	0.222	0.2754
7	CRT-CS	515ms	8	46	8	88	64	92%	15%	0.36	0.5	0.343	0.3153
8	CRT-PLS	531ms	37	19	5	94	84.516	95%	66%	0.36	0.119	0.168	0.3153
9	DA-PLS	452ms	25	21	16	83	74.483	84%	54%	0.267	0.314	0.202	0.2726
10	LDA	593ms	36	20	16	83	76.774	84%	64%	0.267	0.308	0.194	0.2726

Algo -Algorithm names, CTime- Computing Time, TP-True Positive, FN-False Negative, FP-False Positive, TN-True Negative, Acc-Accuracy, Spec-Specificty, Sen-Senitivity, CV Erate-Cross Validation Error rate, P(Prec)-Positive Precision, N(Prec)-Negative Precision, BE rate-Bootstrap Validation Error rate.

SVM, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA in a lowest computing time that we have experimented with a dataset. A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results.

Table 3 shows confusion matrix.

	Classified as Healthy	Classified as not healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

From the confusion matrix to analyze the performance criterion for the classifiers in disease detection accuracy, precision, recall have been computed for all datasets. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted.

Recall is the percentage of positive labelled instances that were predicted as positive. The fitness criteria are calculated as follows:

$$\text{Sensitivity} = TP/(TP+FN)$$
$$\text{Specificity} = TN/(FP+TN)$$
$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

Step 1: The ten algorithms can be filtered by using lowest computing time (<550ms). The ten can be reduced seven algorithms namely (**SVM, PNN, BLR, MLR, CRT, CS-CRT and PLS-DA)**.

S.no							Acc	Spec	Sen	CV	P	N	
	Algo	CTime	TP	FN	FP	TN	%	%	%	Erate	(Prec)	(Prec)	BVErate
1	SVM	546ms	24	30	14	82	70.667	85%	44%	0.29	0.368	0.268	0.2929
2	PNN	546ms	42	12	39	57	66	59%	78%	0.34	0.482	0.174	0.3406
3	BLR	515ms	32	22	19	77	72.667	80%	59%	0.273	0.373	0.222	0.2754
4	MLR	530ms	32	22	19	77	72.667	80%	59%	0.273	0.373	0.222	0.2754
5	CRT - CS	515ms	8	46	8	88	64	92%	15%	0.36	0.5	0.343	0.3153
6	CRT-PLS	531ms	37	19	5	94	84.516	95%	66%	0.36	0.119	0.168	0.3153
7	DA	452ms	25	21	16	83	74.483	84%	54%	0.267	0.314	0.202	0.2726

Step 2: The above algorithms can filtered by using positive precision values. If the precision value is greater than 0.1.we get the six algorithms namely **(SVM, PNN, BLR, MLR, CRT and PLS-DA).**

s.no	Algo	CTime	TP	FN	FP	TN	Acc %	Spec %	San %	CV Erate	P (Prec)	N (Prec)	BVErate
1	SVM	546ms	24	30	14	82	70.6667	85%	44%	0.29	**0.368**	0.2679	0.2929
2	PNN	546ms	42	12	39	57	66	59%	78%	0.34	**0.482**	0.1739	0.3406
3	BLR	515ms	32	22	19	77	72.6667	80%	59%	0.2733	**0.373**	0.222	0.2754
4	MLR	530ms	32	22	19	77	72.6667	80%	59%	0.2733	**0.373**	0.222	0.2754
6	CS-CRT	531ms	37	19	5	94	84.5161	95%	66%	0.36	**0.119**	0.1681	0.3153
7	PLS-DA	452ms	25	21	16	83	74.4828	84%	54%	0.2667	**0.314**	0.2019	0.2726

Step 3: The above algorithms can filter by using Cross Validation Error rate (< 0.3) i.e. lowest error rate. The above six algorithms can be reduced. We get four algorithms namely **(SVM, BLR, MLR, and PLS-LDA)**

s.no	Algo	CTime	TP	FN	FP	TN	Acc %	Spec %	San %	CV Erate	P (Prec)	N (Prec)	BVErate
1	SVM	546ms	24	30	14	82	70.6667	85%	44%	**0.29**	**0.368**	0.2679	0.2929
2	BLR	515ms	32	22	19	77	72.6667	80%	59%	**0.2733**	**0.373**	0.222	0.2754
3	MLR	530ms	32	22	19	77	72.6667	80%	59%	**0.2733**	**0.373**	0.222	0.2754
4	PLS-DA	452ms	25	21	16	83	74.4828	84%	54%	**0.2667**	**0.314**	0.2019	0.2726

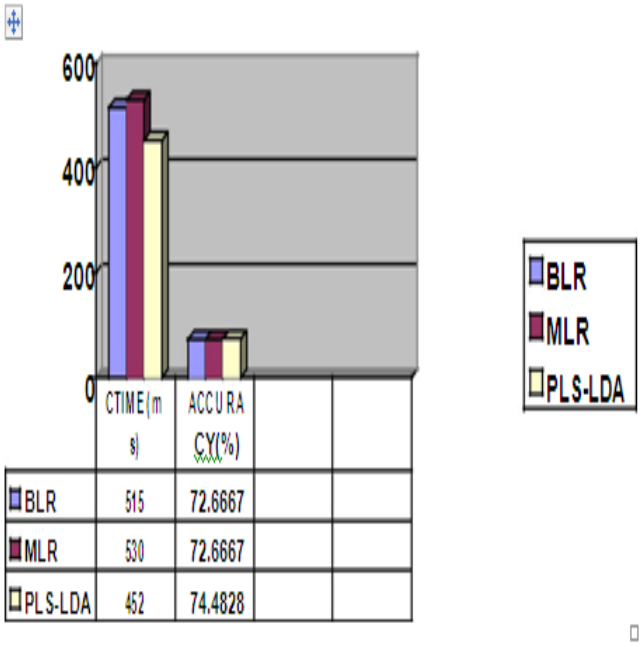
Step 4: The above algorithms can filter by using Bootstrap Validation Error rate (< 0.29) i.e. lowest error rate. The above four algorithms can be reduced. We get three algorithms namely **(PNN, BLR and MLR)**

s.no	Algo	CTime	TP	FN	FP	TN	Acc %	Spec %	San %	CV Erate	P (Prec)	N (Prec)	BVErate
1	BLR	515ms	32	22	19	77	72.6667	80%	59%	**0.2733**	**0.373**	0.222	**0.2754**
2	MLR	530ms	32	22	19	77	72.6667	80%	59%	**0.2733**	**0.373**	0.222	**0.2754**
3	PLS-DA	452ms	25	21	16	83	74.4828	84%	54%	**0.2667**	**0.314**	0.2019	**0.2726**

Step 5: The above algorithms can filter by using highest accuracy and lowest computing time. The above three algorithms can be reduced to one. We get best one for **PLS-DA.**

s.no	Algo	CTime	TP	FN	FP	TN	Acc %	Spec %	San %	CV Erate	P (Prec)	N (Prec)	BVErate
1	PLS-DA	452ms	25	21	16	83	74.4828	84%	54%	**0.2667**	**0.314**	0.2019	**0.2726**

Step 6: Stop the process. We get the best one.



	CTIME(ms)	ACCURACY(%)		
BLR	515	72.6667		
MLR	530	72.6667		
PLS-LDA	452	74.4828		

Figure 2: Predicted Accuracy

The step5 consists of values of different classification. According to these values the accuracy was calculated. The figure 2 represents the resultant values of above classified dataset using data mining supervised classification algorithms and it shows the highest accuracy and lowest computing among the three. It is logical from chart that compared on basis of performance and computing time, precision value, Error rate (10 fold Cross Validation, Bootstrap Validation) and finally the highest accuracy and again lowest computing time. PLS-LDA algorithm shows the superior performance compared to other algorithms.

CONCLUSION

The main goal CPDMA is to get best algorithms that describe given data from multiple aspects. The algorithms are very necessary for intend an automatic classification tools. With help of automatic design tools to reduce a wait in line at the experts. The PLS-LDA was the best one among tens (five criteria are satisfied). Three axis are used the redundancy cut value is 0.0250, positive and negative values are predicted based on the recall and 1-precision values. It can be classified as function as positive and negative and finally constant value of positive and negative. The first one is computing time in 452 milliseconds it is the lowest, second one is Cross Validation error rate is 0.2667 , third positive precision values are greater than 0.1 , fourth one Bootstrap Validation error rate is 0.2726 lowest(i.e. repetition is 1, test error rate 0.2747,Bootstrap ,Bootstrap+) compare to others and finally three values(Accuracy, Specificity and Sensitivity) are calculated by using formula and the prediction one is Accuracy .Then the Accuracy of PLS-LDA is 74% from the above results **PLS-LDA** algorithm plays a vital role.

REFERENCES

[1] Elma kolce(cela),Neki Frasheri "A Literature Review of Data Mining Techniques used in Healthcare Databases". ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-7288.

[2] D.S.Kumar, G.Sathyadevi, S.Sivanesh Decision(2011) ."Support System for Medical Diagnosis Using Data Mining ".

[3] N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems ". A Survey

[4] F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O(2009). "Challenges in Data Mining on Medical Databases ".IGI Global pp. 502-511.

[5] AshaRajkumar,Sophia Reena.G.(2010) "Diagnosis of Heart Disease Using DataMining

Algorithm".Global Journal of Computer Science and Technology-Page 38,Vol-10,Ver-1.0.

[6] E.Knorr.E and R.Ng(1998) "Algorithms forming distance-based outliers in large datasets". In proc.1998 int.Conf.Very Large Data Bases (VldB'98).pages 392-403, New York, NY, Aug.1998.

[7] E.Jiawei Hen and Micheline Kamber(2006) "DataMining Concepts and Techniques ".CA:Elsevier Inc,SanFranciso,2006.

[8] U.M.Piatetsky-Shapiro and G.Smyth (1996) "From Data Mining to Knowledge Discovery : An Overview".pp.1-36, 1996.

[9] S.C.Liao & M.Embrenchts(2000) "DataMining techniques applied to medical information" Med.Inform, pp.81-102,2000

[10] L.Breiman,J.Friedman,J.Olsen C.Stone (1984)"Classification and Regression Trees"Chapman & Hal,1984.

[11] A.Khemphila,V.Boojing(2010) "Comparing Performance of logistic regression,decision tree and neural network for classifying heart disease patients",Proc.of International Conference on Computer Information System and Industrial Management Application,pp.193-198.

[12] K.Srinivas, B.Kavitha Rani,A.Govrdhan(2010),Applications of DataMining Techniques in health care and Prediction Heart Attacks.(IJCSE) International Journal on Computer Science and Engineering vol-02,pp.250-255.

[13] D.Rubben,Jr.Canals(2009) "DataMining in Health care :Current Applications and Issues".

[14] Tanagra DataMining tutorials http:// data-mining-tutorials-blogspot.com.

[15] UCI Machine Learning Repository pima Indian diabetes dataset

[16] Smith, J.,W., Everhart, J.,E., Dickson, W.,C., Knowler, W.,C. and Johannes, R.,S., Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in Proceedings of the Symposium on Computer Applications and Medical Care, IEEE Computer Society Press, 261- 65, 1988.