Classification of Vertebral Column using Naïve Bayes Technique

Sony Krishna Reddy
M.Tech Student
Department of Information Technology
V.R.Siddhartha Engineering College

Sandhya Rani Kodali
M.Tech Student
Department of Information Technology
V.R.Siddhartha Engineering College

Jaya Lakshmi Gundabathina
Assistant Professor
Department of Information Technology
V.R.Siddhartha Engineering College

ABSTRACT
Medical database contain data in various formats like ECG,EEG,X-rays, textual data etc., This data is not located on the same system, it may be distributed amongst various computers depending on data source. This makes medical data retrieval more complex process. So there is need for data mining tools in medical information processing systems to be effective and user friendly. This paper focuses on finding the machine learning methods which can be applied to extract the data useful for medical data analysis and also to patients to search for any relevant information about diseases or analogies.

General Terms
classification algorithms, Bayesian techniques.

Keywords
Naïve, vertebral, classification, hernia

1. INTRODUCTION
Data mining "is extraction of valuable information from large amounts of heterogeneous forma of data [1]" goal is knowledge discovery and demonstration in human readable form. The data comes from different sources like commercial, scientific or government backgrounds. Commercial entities use statistics gathered through data mining techniques to market their products in better way and about customer realtions.Scientific communities use to discover a association between people getting cancer or about location of a nuclear plant. The government use data mining techniques to uncover patterns in their data like to find unusual behavior to prevent terrorist attack.

1.1 Knowledge discovery in databases is a six step process:
- Data warehousing
- Data selection
- Data preprocessing
- Data transformation
- Data mining
- Interpretation/Evaluation

The techniques used in data mining are link analysis(association rules, sequential patterrns,time sequences),predictive modelling(tree induction, neural nets, regression),database segmentation(clustering),deviation detection(visualisation) and classification estimation.

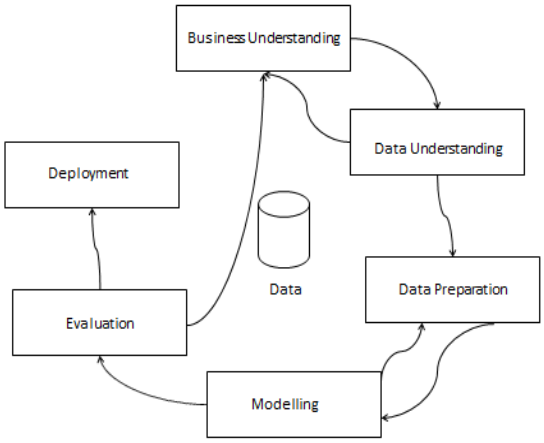

Fig. 1 A knowledge discovery process

Extracting associations in large databases leads to the discovery of useful and previously unknown data. For health care interventions there is a precise need for finding the patterns for purpose such as supervisory management like disease management ,case management[2].Data mining involves the creation of prediction (or classification) models, segmentation (or clustering) records based on similarity of features and discovery of association rules (or patterns).

1.2 Medical Data Mining
Any health care enterprise delivering medical services big challenge is maintenance of medical data i.e., patients identification data, medical record information. The technical staffs of enterprise maintain reliable high-quality data, as low-data quality is a source of medical errors which lead to heavy cost for enterprise [3] Data mining and knowledge discovery is the process of finding patterns, trends and regularities by examining through large amounts of data [4]. The importance of Medical Data Mining (MDM) is to help the physician to make the final decision without hesitation, minimizing diagnostic errors refining diagnostic speed and increasing the quality of medical treatment. [2].In this study we review MDM from different viewpoints. We start with

highlighting the special characteristics of medical data and talk over the requirements of data mining systems to cope with medical data problems and difficulties. We present a periodical of some of those proposed methods in the medical domain to show what the dissimilar techniques are and methods which have been applied to medical data. In the past, several statistical methods have been used for modeling in the area of disease diagnosis. These methods necessitate prior molds and are less capable of dealing with massive and complex nonlinear and dependent data. However, data mining has proven to be more powerful and effective and it provides procedures for discovering useful. [2]

2. DATA MINING CLASSIFICATION METHODS

The data mining contains of various methods. Different methods serve different purposes, each method proposing its own advantages and disadvantages. However, most data mining methods normally used for this review are of classification grouping as the applied prediction techniques allocate patients to either a ”gentle” group that is non-cancerous or a ”malicious” group that is cancerous and generate rules for the same. Hence, the breast cancer diagnostic problems are essentially in the scope of the widely discussed classification problems.

In data mining, classification is one of the most key tasks. It maps the data in to predefined goals. It is a supervised learning as goals are predefined. The aim of the classification is to build a classifier based on specific cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the field based on the standards of other attributes. The commonly used methods for data mining classification tasks can be categorized into the following: [3]

2.1 Naïve Bayes Technique

The Naïve Bayes Classifier technique is mainly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outclass more refined classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state

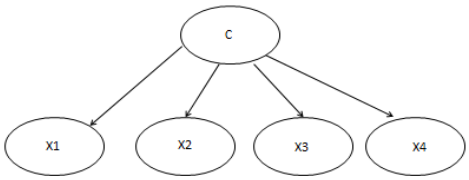

Fig. 2 A Naïve Bayes Network

Why chosen Naïve Bayes: Naive Bayes or Bayes' Rule is the foundation for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of discovering and considerate data.

Bayes Rule: A conditional probability is the likelihood of specific decision, C, given some evidence/observation, E, where a need relationship occurs between C and E.[6]

This probability is denoted as P(C |E) where

$$P(C \mid E) = \frac{P(E \mid C)\, P(C)}{P(E)}$$

The Naïve Bayesian Classification: The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

Let D be a training set of tuples and their associated class labels. As usual, each tuple is denoted by an n-dimensional attribute vector, X=(x 1, x2,…, xn), illustrating n measurements made on the tuple from n attributes, respectively, A1, A2,.., An. [6]

Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class taking the highest posterior probability, trained on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

$$P(Ci|X) > P(Cj|X) \quad \text{for } 1 \le j \le m,\ j \ne i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem [6]

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

As P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is generally expected that the classes are equally likely, that is, P(C1)=P(C2) =…=P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the numeral of training tuples of class Ci in D. [6]

Given data sets with many attributes, it would be particularly computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional objectivity is made. This supposes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X \mid Ci) = \prod_{k=1}^{n} P(xk \mid Ci)[5]$$

$$=P(x1|Ci)x\ P(x2|Ci)x\ldots\ P(xm|Ci).$$

We can simply estimate the probabilities P (x1|Ci), P(x2|C i),… ,P(xm|Ci) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X. For each attribute, we aspect at whether the attribute is categorical or continuous-valued. For instance, to compute P(X|Ci), we consider the following:

(a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci,D|, the number of tuples of class Ci in D.[6]

(b) If Ak is continuous valued, then we need to do a bit more work, but the design is pretty direct. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

We need to figure μci and σci, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci. We then pad these two quantities into the

above equation.5. In order to predict the class label of X, P(X|Ci)P(C i) is evaluated for each class Ci. The classifier calculates that the class label of tuple X is the class Ci if and only if

$$P(X|Ci)P(Ci)>P(X|Cj)P(Cj)$$
$$\text{for } 1 \leq j \leq m, j \neq i[5]$$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the determined.

However, between the different approaches and techniques used for medical uses, in this paper we are concerned with the use of Naïve Bayes (NB) for medical classification. In the following we discuss its simple types and how it outfits for this domain.

Naïve Bayesian classifier, or simply naïve bayes (NB), is one of the most effective and efficient classification algorithms. It is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions.

Given a set of training occurrences with class labels and a test case E represented by n attribute values (a1, a2... an), Bayesian classifiers use the following equation to classify E:

$$C\,NB(E)=\arg\,cmax\; p(c)\prod i=1p(ai/c):$$

where, cNB (E) denotes the classification given by NB on test case E. [4]

Although objectivity is usually a poor assumption, in practice NB often enters well with much more refined techniques. In a large-scale comparison of naïve Bayes classifier with state-of-the-art algorithms for decision tree induction, instance-based learning and rule induction, conducted by on standard datasets NB is superior to the other learning schemes, even on datasets with substantial feature dependencies.

A variety of revisions to NB in the works have been studied in order to improve upon its good performance while preserving its efficiency and easiness. NB has proven its effective application, often reported as "amazingly" correct, in text classification, medical diagnosis and systems performance management however, as mentioned previously in this paper we are concerned on its application to medical data and how it handles the different problems in this domain. In the following, based on the discussed requirements of medical data mining systems, we see how this approach is applicable for mining medical data.[4]

3. MEDICAL DATA MINING WITH NAÏVE BAYES

NB as a benchmark algorithm that in any medical domain has to be tried before any other advanced method. The simple methods are better in medical data mining and this makes NB performs well for such data. Associated to other classifiers, NB is simple, computationally efficient, requires reasonably little data for training do not have lot of parameters and is certainly robust to missing and noise data One of the main advantages of NB approach which is interesting to physicians, is that all the available information is used to describe the conclusion. This explanation seems to be "regular" for medical diagnosis and prediction i.e. is nearby to the way how physicians diagnose patients [7]

When allocating with medical data, naïve bayes classifier takes into description evidence from many attributes to make the final prediction and provides clear descriptions of its decisions and thus it is considered as one of the most useful classifiers to support physicians' decisions.

Successful applications of NB to medical data have been conveyed by various researchers in the literature. NB with six algorithms (Assistant-R, Assistant-I, LFC, back propagation, k-NN and semi-NB). The result was that NB classifier outclassed all the algorithms on five out of eight medical diagnostic problems.

However, even with small data sets, naïve bayes have shown that it can hypothesis practically exact prognostic models as verified by, who used naïve bayes classifier with a data set which includes only 68 patients. In a comparative study of discretization methods for medical data mining, it suggests that on an average the NB classifier with MDL discretization seems to be the best performer compared to popular variants of NB and non-NB classifiers (such as DT, k-NN and LR).[8]

4. EXPERIMENT SETUP

The problem area of vertebral column in humans is spinal cord. Here we have taken data set having values for six biomechanical features used to classify orthopaedic patients into 3 classes (normal, disk hernia or spondilolysthesis) or 2 classes (normal or abnormal). [9]The first task involves in categorizing patients as fit in to one out of three categories: Normal (100 patients), Disk Hernia (60 patients) or Spondylolisthesis (150 patients). For the second task, the categories Disk Hernia and Spondylolisthesis were combined into a single category labelled as 'abnormal'. Thus, the second task consists in classifying patients as fitting to one out of two categories: Normal (100 patients) or Abnormal (210 patients). Each patient is characterized in the data set by six biomechanical attributes resultant from the shape and alignment of the pelvis and lumbar spine (in this order): pelvic occurrence, pelvic tilt, lumbar lordosis position, sacral slope, pelvic radius and mark of Spondylolisthesis. The following settlement is used for the class labels: DH (Disk Hernia), Spondylolisthesis (SL), Normal (NO) and Abnormal (AB). A herniated disk and Spondylolisthesis are two possibly painful circumstances that can affect the steadiness and function of the spinal column. While herniation affects the discs between the spinal bones (vertebrae), Spondylolisthesis affect the bones themselves.

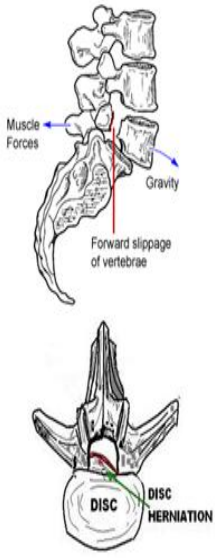

Fig. 3 Disk Hernia and Spondilolysthesis

The attributes in data set are 7:

Table 1 : Classifier Model for Full Training set

	Class	
Attribute	**Abnormal**	**Normal**
	(0.68)	(0.32)
pelvic_incidence		
mean	64.69	51.6786
std. dev.	17.6236	12.3124
weight sum	210	100
precision	0.3356	0.3356
pelvic_tilt		
mean	19.7882	12.8154
std. dev.	10.4901	6.7344
weight sum	210	100
precision	0.1812	0.1812
lumbar_lordosis_angle		
mean	55.9246	43.5395
std. dev.	19.6145	12.2925
weight sum	210	100
precision	0.4005	0.4005
sacral_slope		
mean	44.9213	3 8.8601
std. dev.	14.4784	9.5732
weight sum	210	100
precision	0.3859	0.3859
pelvic_radius		
mean	115.0741	123.8883
std. dev.	14.0673	8.9667
weight sum	210	100
precision	0.3009	0.3009
degree_spondylolisthesis		
mean	37.7896	2.2384
std. dev.	40.5897	6.2791
weight sum	210	100
precision	1.3903	1.3903

Table 2 :Classifier Model for Full Training set

Attribute	**Hernia**	**Spondylolisthesis**	**Normal**
	(0.19)	(0.48)	(0.32)
pelvic_incidence			
mean	47.6375	71.511	51.6786
std. dev.	10.6179	15.0629	12.3124
weight sum	60	150	100
precision	0.3356	0.3356	0.3356
pelvic_tilt			
mean	17.397	20.7447	12.8154
std. dev.	6.9559	11.4675	6.7344
weight sum	60	150	100
precision	0.1812	0.1812	0.1812
lumbar_lordosis_angle			
mean	35.4719	64.1057	43.5395
std. dev.	9.6615	16.341	12.2925
weight sum	60	150	100
precision	0.4005	0.4005	0.4005

sacral_slope			
mean	30.2575	50.7869	38.8601
std. dev.	7.5055	12.269	9.5732
weight sum	60	150	100
precision	0.3859	0.3859	0.3859
pelvic_radius			
mean	116.4663	114.5173	123.8883
std. dev.	9.2839	15.5397	8.9667
weight sum	60	150	100
precision	0.3009	0.3009	0.3009
degree_spondylolisthesis			
mean	2.5025	51.9044	2.2384
std. dev.	5.56	39.9608	6.2791
weight sum	60	150	100
precision	1.3903	1.3903	1.3903

Table 3: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.733	0.12	0.928	0.733	0.819	0.575	0.886	0.95	Abnormal
0.88	0.267	0.611	0.88	0.721	0.575	0.886	0.757	Normal
Weighted Avg.	0.781	0.167	0.826	0.781	0.788	0.575	0.886	0.887

Table 4: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.717	0.092	0.652	0.717	0.683	0.603	0.922	0.703	Hernia
0.973	0.088	0.913	0.973	0.942	0.886	0.99	0.989	Spondylolisthesis
0.69	0.071	0.821	0.69	0.75	0.651	0.919	0.855	Normal
Weighted Avg	0.832	0.083	0.833	0.832	0.83	0.755	0.954	0.89

Table 5: Performance study of the algorithm

Accuracy : 83.7419 %
Time taken to build model: 0.1 seconds

5. CONCLUSIONS

Machine Learning (ML) study have been successfully applied to medical data to discover valuable and new knowledge This study revised existing state of medical data mining ,and classification techniques that we can relate for medical data. Naive Bayes classification approach has been discussed and its main features which gives better results for medical data mining, relating Naive Bayes classification technique for certain medical data set to classify the data. The significance of Medical Data Mining (MDM) is to support the physician to make the final decision without disinclination, underestimating analytical errors improving diagnostic speed and increasing the eminence of medical treatment.

6. REFERENCES

[1] Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han

[2] Tatiana Semenova,Warwick Graco,Markus Hegland,Graham Williams, " Effectiveness of mining association rules for identifying trensds in alrge health databases ",2001 - users.csc.calpoly.edu

[3] Rippen HE, Yasnoff WA. Building the National Health Information InfrastructureJ AHIMA. 2004 May;75(5):20-6.

[4] A. Bakar and Z. Othman," Asian Network for Scientific Inforrnation Medical Data Classification with Naïve Bayes Approach Al-Aidaroos ",Information Technology Journal 11(9):1166-1174© 2012

[5] MIR Labs http://www.ijcir.comEffective Discretization and Hybrid featureselection using Naïve Bayesian classifier forMedical datamining (2002)

[6] G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)Decision Support in Heart Disease Prediction System usingNaive

[7] Al-Aidaroos, K.M., A.A. Bakar and Z. Othman, 2010. Naive Bayes variants in classification learning. Proceeding of the International conference on Information Retrieval and Knowledge Management (CAMP 2010), March 17-18, 2010, Shah Alam, Selangor, pp: 276-281.

[8] Kononenko, I., 2001. Machine learning for medical diagnosis: History, state of the art and perspective. Artif. Intell. Med., 23: 89-109. PubMed Frank, A. & Asuncion, A. (2010).

[9] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.