SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES

K.Srinivas[1], Dr. G.Raghavendra Rao[2] and
Dr. A.Govardhan[3]

[1]Department of Computer Science and Engineering JITS, Karimnagar, AP, India.
jaya_konda@yahoo.com
[2]Principal, NIE Institute of Technology & Science, Mysore, Karnataka, India
princi_nieit@nie.ac.in
[3]Principal , JNTUH College of Engineering, Kondagattu, , AP, India.
govardhan_cse@yahoo.co.in

ABSTRACT

Data mining is the non trivial extraction of implicit, previously unknown and potentially useful information from data. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. This paper presents about the various existing techniques, the issues and challenges associated with them. The discovered knowledge can be used by the healthcare administrators to improve the quality of service and also used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. In this paper we discuss the popular data mining techniques namely, Decision Trees, Naïve Bayes and Neural Network that are used for prediction of disease.

KEYWORDS

KDD, data mining, machine learning algorithms, classifiers, disease prediction, time series, ARIMA, DSDM.

1. INTRODUCTION

With the ever-growing complexity in recent years, huge amounts of information in the area of medicine have been saved every day in different electronic forms such as Electronic Health Records (EHRs) and registers. These data are collected and used for different purposes. Data stored in registers are used mainly for monitoring and analyzing health and social conditions in the population. The unique personal identification number of every inhabitant enables linkage of exposure and outcome data spanning several decades and obtained from different sources. The existence of accurate epidemiological registers a basic prerequisite for monitoring and analyzing health and social conditions in the population. Some registers are state-wide, cover the whole collieries population, and have been collecting data for decades. They are frequently used for research, evaluation, planning and other purposes by a variety of users in terms of analyzing and predicting the health status of individuals.

1.1. Focus of this Survey

Recent studies have documented poor population health outcomes in coal mining areas. These findings include higher chronic cardiovascular disease (CVD) mortality rates [Hendryx, 2009] and higher rates of self-reported CVD [Hendryx and Ahern, 2008]. The risk for CVD is influenced by environmental, behavioral, genetic, demographic, and health services variables.

Risk behaviors, in turn, are related to lower socio economic status (SES); low SES persons are more likely to smoke, consume poor quality diets, and engage in sedentary lifestyles. Coal mining areas are characterized by lower SES relative to non-mining areas, suggestive of higher CVD risk. Environmental agents that contribute to CVD include arsenic, cadmium and other metals, non-specific particulate matter (PM), and polycyclic aromatic hydrocarbons (PAHs). All of these agents are present in coal or introduced into local ambient environments via activities of coal extraction and processing. Most previous research on population health in coal mining areas has employed state-level mortality data rather than individual-level data. An exception was a study of self-reported chronic illness in relation to coal mining; this study was limited to a non-standard assessment instrument with limited individual-level covariates in Singareni Collieries, Andhra Pradesh state in country India. The current study uses National Behavioral Risk Factor Surveillance System (BRFSS) data to assess CVD risk in coal mining areas before and after control for individual-level covariates including smoking, obesity, co-morbid diabetes, alcohol consumption and others. We test the hypothesis that CVD rates will be significantly elevated for residents of coal mining regions after controlling for covariates, suggestive of an environmental impact.

1.2. Data Mining concepts in Health care

Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. It is a process that is developed to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. In practice, the two primary goals of data mining tend to be *prediction* and *description. Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description,* on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories.

- *Predictive data mining*: Predictive models can be used to forecast explicit values, based on patterns determined from known results. For example, from a database of customers who have already responded to a particular offer, a model can be built that predicts which prospects are likeliest to respond to the same offer.
- *Descriptive data mining*: Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

2. DATA PREPARATION

High quality data is a prerequisite for any data mining technique. Prior to data modeling, the data needs to be prepared. The objective at this stage is two-fold; to obtain data prepared in the form required by the data mining algorithms and to expose as much information as possible for data modeling. There are four types of data preprocessing techniques to transform data and

improve its quality to suit the requirements of the learning algorithms [Grabmeier & Rudolph] which are first is data cleansing. It is applied to the data by filling in missing values, removing or smoothing noise, identifying or removing outliers and inconsistencies in the data. The second is data integration. It merges the data from multiple sources into coherent data stores. The third is data transformation. It transforms and consolidates data into forms appropriate for mining, and the forth is data reduction. It reduces the data size by aggregation, elimination of redundant features, or clustering.

2.1. Data Preprocessing

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks.

2.2 Outlier detection and removal

Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such non representative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

a) Detect and eventually remove outliers as a part of the preprocessing phase, or
b) Develop robust modeling methods that are insensitive to outliers.

2.3. Scaling, encoding and selecting features

Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [-100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data mining process. Data preprocessing steps should not be considered completely independent from other data-mining phases. In very iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a prior knowledge in the form of application specific scaling and encoding.

Before the data undergo data mining, they must be prepared in a pre-processing step that removes or reduces noise and handles missing values. Relevance analyses for omitting unnecessary and redundant data, as well as data transformation, are needed for generalizing the data to higher-level concepts. Techniques such as Expectation Maximization (EM) method are used for replacing the missing values. Various algorithms such as

- Parameter estimation method that falls within the general framework of maximum likelihood estimation and is an iterative optimization algorithm.
- Hotelling (1936) developed Canonical Correlation Analysis (CCA) as a method for evaluating linear correlation between sets of variables. The method allows investigation of the relationship between two sets of variables and identification of the important

ones. It can be used as a dimension reduction technique to preserve the character of the original data stored in the registers by omitting data that are nonessential. The objective is to find a subset of variables with predictive performance comparable to the full set of variables.

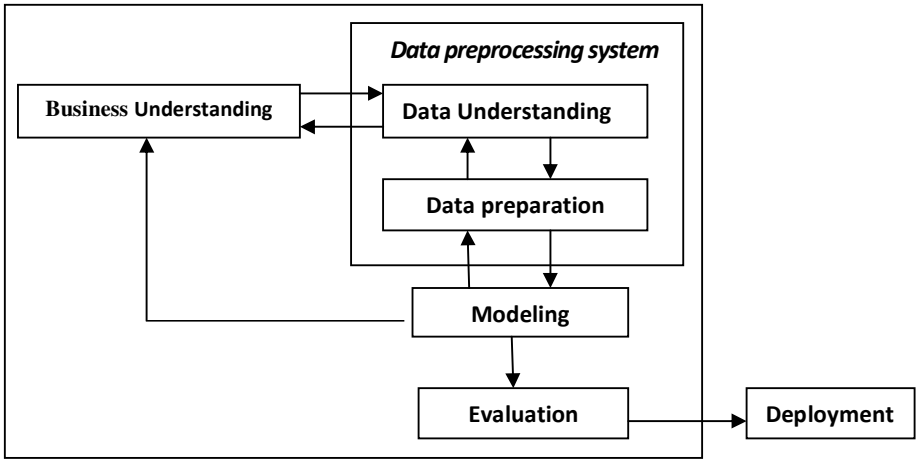

Figure 1. Data preprocessing system showing the steps in processing

Data pre-processing, including feature extraction for statistical information extraction and classification tasks, is a considerable bottleneck. Existing advanced multi-pattern matching algorithms are crucial in time-sensitive processing of high volume data.

The explosion in the size and number of information sources observed in recent years emphasizes the need for identifying large concepts sets in vast amounts of data. In most cases, fast parallel multi-pattern matching algorithms are needed for rapid and timely processing of data. In the news domain, in order to process large corpora, resource-driven information extraction relies on matching large free text databases against massive concept sets drawn from resources such as gazetteers, ontologies, semantic networks etc. This processing step is often time consuming, yet inherently parallelizable.

The pre-processing system shown in figure 1 shall support data understanding and modification by providing mutual interoperability of all the tools. Using advanced visualization modules, the user will better navigate in graphs. Interoperability of the visualization and the modification tool will enable the user to set-up parameters of a filter, which will be used in the modification phase, directly within an interactive histogram provided by the visualization tool. Such interoperability significantly speeds up the pre-processing procedure and makes it much more interesting and valuable.

2.4. Preprocessing Techniques

There are various stages in data mining for the analysis of data. Appropriate tools and techniques for the process are required in order to obtain useful information. In this particular study, the focus will be on the stages of pre-processing and application of the data mining algorithm.

2.4.1. Wavelet Transformation

The first part of the research deals with applying the wavelet transform as the technique to pre-process the large amount of time series medical data. Time series data are series of values of some feature of an event that are recorded over a period of time. These clinical data are inconsistent and contains noise. Therefore, wavelet transform is applied. It plays an important role in data mining due to its easy accessibility and practical use. Wavelet transformation provides efficient and effective solutions to many data mining problems. The wide variety of wavelet-related methods has been applied to a wide range of data mining problem. Wavelet transformation is the tool that divides up data, function, or operators into different frequency components with a resolution that matches to its scale. It provides economical and informative mathematical representation of many objects of interest. The tool is used in order to observe its ability to provide presentations of data that make mining process more efficient and accurate. After pre-processing, the clinical data goes through the data mining algorithm in order to obtain the desired representation of the information. There are number of data mining algorithms. Applying appropriate data mining algorithms to the prepared data is very important in the data mining process. The selection on the type of algorithm to use depends on the given data mining problem.

2.4.2. k-means Clustering Algorithm

The second part of this study deals with the clustering algorithm, specifically k-means clustering algorithm. Data clustering is one of the most well known and commonly used techniques in data mining. It is used to explore irregularities in the data distribution. K-means algorithm is chosen due to its simplicity and speed in analyzing large amount of data. Comparison is made by applying the k-means algorithm to the original data and data that are pre-processed by wavelet transform to prove that pre-processed data give better overall results.

2.5. Dimensionality Reduction and Feature Selection

When processing large databases, one faces two major obstacles such as numerous samples and high dimensionality of the feature space. For example, the documents are represented by several thousands of words; images are composed of millions of pixels, where each word or pixel is here understood as a feature. Currently, processing abilities are often not able to handle such high dimensional data, mostly due to numerical difficulties in processing, requirements in storage and transmission within a reasonable time. To reduce the computational time, it is common practice to project the data onto a smaller, latent space.

Moreover, such a space is often beneficial for further investigation due to noise reduction and desired feature extraction properties. Smaller dimensions are also advantageous when visualizing and analyzing the data. Thus, in order to extract desirable information, dimensionality reduction methods are often applied. The overall idea is to determine the coordinate system where the mapping will create low-dimensional compact representation of the data whilst maximizing the information contained within.

There are many solutions to this problem. Several techniques for dimensionality reduction have been developed which use both linear and non-linear mappings. Among them are, for example, low-dimensional projections of the data, neural networks self-organizing maps. One can apply second order methods which use the covariance structure in determining directions. Principal Component Analysis that restricts directions to those are orthogonal. Factor Analysis which additionally allows the noise level to differ along the directions and Independent Component Analysis for which the directions are independent but not necessarily orthogonal.

2.5.1. Random Projection

A lot of research studies have focused on an investigation of the Random projection method. The method is simple and it does not require the presence of the data whilst creating the projection matrix. Many find this a significant advantage. Random projection is based on the set of vectors which are orthogonalized from the normal distributed random matrix. The orthogonal projection vectors ($R.R^T = I$) preserves the distances among the data points. In this way it is possible to project the data to the smaller dimensional space while preserving fair separation of the clusters. Note however, that R is orthogonal only column-wise, i.e. the identity outcome of the dot product of the projection vectors holds only for quadratic R (projecting onto the space of the same dimension as the original).

In the case of projecting onto smaller dimensions, the distance is no longer preserved in an ideal way. Let the D-dimensional data matrix be defined as $X_{D \times N}$, where N is the number of samples. Data is projected with the help of the projection matrix R to K dimensions ($K \leq D$) and the projected data matrix is denoted as $\hat{X}_{KxN}$. Figure 2 presents the algorithm of the Random Projection method.

The Random Projection Algorithm

1. Generate random matrix $R_{KxD} = \{ r_{kd}, k = 1 \ldots K, d = 1 \ldots D\}$ drawn from normal distribution of zero mean and spherical covariance structure $r_{kd} \in N(0, 1)$, $K \leq D$.
2. Orthogonalize each of the projection vectors (use for example Gram- Schmidt algorithm [Steven J. Leon.].
3. Project data on the new set of directions $\hat{X}_{KxN} = R_{KxD} \cdot X_{DxN}$

Figure 2. The Random Projection algorithm

2.5.2. Principal Component Analysis

Principal Component Analysis (PCA) is probably the most widely used technique for dimensionality reduction. It is similar to random projection through performing a linear mapping that uses the orthogonal projection vectors, but in case of PCA the projections are found by diagonalizing the data covariance structure, i.e. the variance along new directions is maximized. Thus, the projection of vector x on the new latent space appointed by orthogonal projection vectors u can be written as $\hat{X}k = u_k^T x$, where $k = 1. . .K$ and $K \leq D$. It is proofed, that the minimum projection error (in LS sense) is achieved when the basis vectors satisfy the eigen-equations $\Sigma.u_k = \lambda_k u_k$ [C. M. Bishop.]. Therefore, the singular value decomposition (SVD) is often performed to find the orthogonal projections. The algorithm for PCA is shown in figure 3. Such a method for determining the latent space is applied by Latent Semantic Analysis (LSA). The most significant disadvantage of the PCA technique is complexity and high memory usage

Principal Component Analysis Algorithm

1. Create the data matrix X

2. Determine covariance $\Sigma = \frac{1}{N} \Sigma_n (xn - \mu)(xn - \mu)T$

3. Determine eigen values λ_k and eigen vectors u_k of the covariance structure. Since Σ is positive and symmetric, λ is positive and real satisfying $\Sigma \cdot u_k = \lambda_k u_k$. λ is found in the optimization process of the characteristic equation $|\Sigma - \lambda I| = 0$.

4. Sort eigen values and corresponding eigenvectors in the descending order.

5. Select $K < D$ and project the data on selected directions:

$\hat{X}_{KxN} = U^T_{DxK} \cdot X_{DxN}$

Figure 3. The Principal Component Analysis algorithm.

The Empirical study on data reduction by using random projection and principal component analysis done by [Sampath Deegalla, Henrik Bostr] considering five image data sets and five micro array data sets are representing two types of high dimensionality classification tasks. The image data sets consist of two medical image datasets (IRMA, MIAS), two object recognition data sets (COIL-100 , ZuBuD) and a texture analysis data set (Outex - *TC 00013*). The IRMA (Image Retrieval and Medical Application) data set contains radiography images of 57 classes, where the quality of the images varies significantly. The COIL-100 (Columbia university image library) data set consists of images of 100 objects, while ZuBuD (Zurich Building Image Database) contains images of 201 buildings in Zurich city. MIAS (The Mammography Image Analysis Society) mini mammography database contains mammography images of 7 categories and finally Outex (University of Oulu Texture Database) image data set contains images of 68 general textures. The five micro array data sets are: Leukemia, Colon Tumor, Central Nervous, Srbct (small, round, blue, cell tumors) and Lymphoma.

Experiment are performed for the dimensionality reduction by considering all image data sets, color images have been converted into gray scale images and then resized into 32 × 32 pixel sized images, and where the brightness values are the only considered features. Therefore, all image data sets contain 1024 attributes. The number of instances and attributes for all data sets are shown in Table 1.

MATLAB has been used to transform the original matrices into projected matrices using PCA, through the singular value decomposition (SVD) implementation of PCA. The Waikato Environment for Knowledge Analysis (WEKA) has been used for RP as well as for the nearest neighbor classifier. The accuracies were estimated using tenfold cross validation, and the results for RP is the average from 30 runs to account for its random nature.

Table 1. Description of data.

Data set	Instances	Attributes	# of Classes
IRMA	9000	1024	57
COIL100	7200	1024	100
ZuBuD	1005	1024	201
MIAS	322	1024	7
Outex	680	1024	68
Colon Tumor	62	2000	2
Leukemia	38	7129	2
Central Nervous	60	7129	2
Srbct	63	2308	4
Lymphoma	62	4026	3

The experimental results show that reducing the dimensionality using PCA results higher accuracy for most of the data sets. In Table 2, it can be seen that only a few principal components is required for achieving the highest accuracy which is evident in figure 5. However, RP typically requires a larger number of dimensions compared to PCA to obtain a high accuracy. Classification accuracy using PCA typically has its peak for a small number of dimensions, after which the accuracy degrades. In contrast to this, the accuracy of RP generally increases with the number of dimensions. Hence, this shows that PCA is more sensitive to the choice of the number of reduced dimensions than RP. However, for all the data sets used in this study, the maximum accuracy obtained by using PCA is higher than the maximum accuracy obtained by using RP. This means that one can expect PCA to be more effective than RP if the number of dimensions is carefully chosen. The experiments also show as in figure 4 that the use of PCA and RP may even outperform using the non-reduced feature set (in 9 respectively 6 cases out of 10).

Table 2. Highest prediction accuracy obtained by nearest neighbor classifier with dimensionality reduction methods (no. of dimensions in parentheses).

Data set	RP		PCA		Original
IRMA	67.01	(250)	**75.30**	(40)	68.29
COIL100	98.79	(250)	98.90	(30)	**98.92**
ZuBuD	54.01	(250)	**69.46**	(20)	59.80
MIAS	44.05	(5)	**53.76**	(250)	43.17
Outex	21.04	(15)	**29.12**	(10)	19.85
Colon Tumor	80.22	(150,200)	**83.05**	(10)	77.42
Leukemia	91.32	(150)	**92.83**	(10)	89.47
Central Nervous	58.22	(150)	**66.33**	(50)	56.67
Srbct	93.23	(200)	**96.45**	(10)	87.30
Lymphoma	97.80	(250)	**99.86**	(20)	98.38

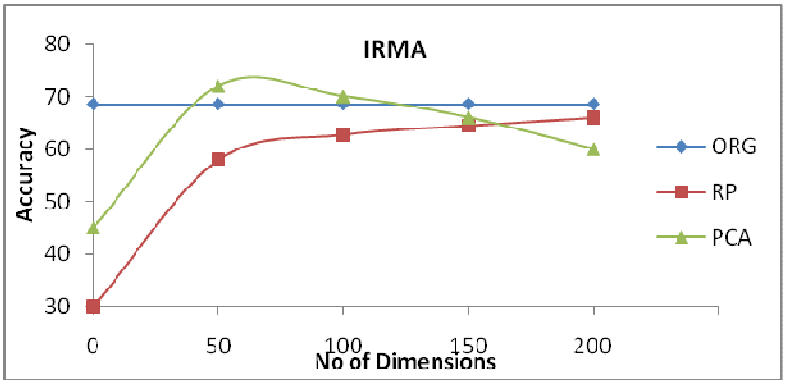

Figure 4. Comparison of the accuracies between Original, PCA and RP based attributes.

When the projected dimension is large enough, the distortion of the distance is not significant but then the time spent for creating and diagonalizing projection matrix is considerable. To summarize, for large dimensional data sets it is not necessarily efficient and satisfying enough to use random projection without significant separability loss, especially when the dimensionality the data is projected onto is high.

2.5.3 Non-negative Matrix Factorization

Some medical data such as images or texts contain only positive values. This information can be utilized by adding the positivity constraint in the optimization process for finding the projection vectors. One possible choice is the Non-negative Matrix Factorization (NMF). The technique is closely related with proposed by Hofmann, Probabilistic Latent semantic Analysis (PLSA) and Saul and Peveira in aggregated Markov model. NMF is based on the decomposition of the data matrix X into two matrix factors W and H so that $X_{D \times N} \approx W_{D \times K} H_{K \times N}$. Both W and H are constrained to be positive, i.e. $W \geq 0$ and $H \geq 0$. The proposed optimization is done in an iterative way minimizing one of the two suggested objective functions

Euclidean distance: $$||X - WH||_2^2 = \sum\sum(x_{dn} - \sum w_{dk} h_{kn})^2$$

or Kullback-Leibler (KL) divergence :

$$D(X \,||\, WH = \sum_d \sum_n (x_{dn} . log \frac{x_{dn}}{\sum_k w_{dk} h_{kn}} - x_{dn} + \sum_k w_{dk} h_{kn})$$

between X and WH. The update rules in case of the Euclidean distance are shown in matrix notation by the equations

$$H^{new} = H^{old}\frac{W^T X}{W^T WH} \qquad W^{new} = W^{old}\frac{XH^T}{WHH^T}$$

Regarding KL divergence, the update rules are taking the form

$$H^{new} = H^{old}\frac{W^T X/WH}{1_{NxK}W^T} \qquad W^{new} = W^{old}\frac{X.WH.H^T}{1_{DxN}H^T}$$

The algorithm is proofed to converge and the modified NMF updates rules are used in segmentation by the aggregated Markov model.

2.6. Data Design

The disease prediction empirically heart attack prediction system uses attributes with dichotomous self-reported measures assessing whether respondents were ever diagnosed with morbidity. Other than these attributes we use regular medical attribute which are used for prediction of heart attack.

Figure 5 is listed with 15 attributes which specify the normal ranges of the diagnosis. Other than regular attributes we use (1) chest pain, (2) heart attack or stroke as additional attributes. A fourth general CVD category measured whether respondents reported the presence of any of these three CVD types. The morbidity categories are self-reported so an exact correspondence to diagnostic categories is uncertain. Covariates included smoking, coded as a three-level variable: current, former (smoked at any time in the past), or lifetime non-smoker. Self-report body mass index (BMI) was coded into: underweight (BMIb18.5); normal (BMI 18.5 to b25); overweight (BMI 25 to b30); and obese (BMI 30 or greater) with the normal weight category serving as the referent.

Predictable attribute
1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))
Key attribute
1. PatientID – Patient's identification number
Input attributes
1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping;value 2: flat;value 3: downsloping)
7. CA – number of major vessels colored by floursopy value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Figure 5. Description of health data attributes

Description of attributes alcohol consumption was coded as average number of drinks per day and was categorized into non-drinker, light drinker (1 or fewer drinks per day), moderate drinker (more than 1 but less than 4), or heavy drinker (4 or more per day). Light drinkers were used as the referent category. Age was coded in years and ranged from 18 to 99. Diabetes co-morbidity was coded yes/no based on the respondent reporting ever being diagnosed with diabetes. Marital status was dichotomized as married or cohabitating versus any other status. Education was scored 1 to 6, ranging from "never attended school or only kindergarten" to "college 4 years or more (college graduate)"; a score of 4 was equivalent to a high school graduate

3. DATAMINING TECHNIQUES IN HEALTH CARE

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques with illustrations of their applications to healthcare. We employed four types of classification algorithms: Bayesian model, neural networks, SVM and decision trees. These models were jigged for inclusion in this research due to their popularity in the recently published documents. The following is a brief introduction to the four classification algorithms and the parameter setting of each model.

3.1 Decision trees

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2, . . . according to the outcome for each case, and apply the same procedure recursively to each subset.

The initial tree is then pruned to avoid overfitting. The pruning algorithm is based on a pessimistic estimate of the error rate associated with a set of *N* cases, *E* of which do not belong to the most frequent class. Instead of *E/N*, C4.5 determines the upper limit of the binomial probability when *E* events have been observed in *N* trials, using a user-specified confidence whose default value is 0.25.

Pruning is carried out from the leaves to the root. The estimated error at a leaf with *N* cases and *E* errors is *N* times the pessimistic error rate as above. For a subtree, C4.5 adds the estimated errors of the branches and compares this to the estimated error if the subtree is replaced by a leaf; if the latter is no higher than the former, the subtree is pruned. Similarly, C4.5 checks the estimated error if the subtree is replaced by one of its branches and when this appears beneficial the tree is modified accordingly. The pruning process is completed in one pass through the tree.
C4.5's tree-construction algorithm differs in several respects from CART (Correlation and Regression Tree) [Breiman L, Friedman] , for instance.

- Tests in CART are always binary, but C4.5 allows two or more outcomes.
- CART uses the Gini diversity index to rank tests, whereas C4.5 uses information-based criteria.
- CART prunes trees using a cost-complexity model whose parameters are estimated by cross-validation; C4.5 uses a single-pass algorithm derived from binomial confidence limits.
- This brief discussion has not mentioned what happens when some of a case's values are unknown. CART looks for surrogate tests that approximate the outcomes when the tested attribute has an unknown value, but C4.5 apportions the case probabilistically among the outcomes.

3.1.1. Ruleset classifiers

Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form "if A and B and C and ... then class X", where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied, the case is assigned to a default class. C4.5 rulesets are formed from the initial (unpruned) decision tree. Each path from the root of the tree to a leaf becomes a prototype rule whose conditions are the outcomes along the path and whose class is the label of the leaf. This rule is then simplified by determining the effect of discarding each condition in turn. Dropping a condition may increase the number *N* of cases covered by the rule, and also the number *E* of cases that do not belong to the class nominated by the rule, and may lower the pessimistic error rate determined as above. A hill-climbing algorithm is used to drop conditions until the lowest pessimistic error rate is found. To complete the process, a subset of simplified rules is selected for each class in turn. These class subsets are ordered to minimize the error on the training cases and a default class is chosen. The final ruleset usually has far fewer rules than the number of leaves on the pruned decision tree.

The principal disadvantage of C4.5's rulesets is the amount of CPU time and memory that they require. In one experiment, samples ranging from 10,000 to 100,000 cases were drawn from a large dataset. For decision trees, moving from 10 to 100K cases increased CPU time on a PC from 1.4 to 61 s, a factor of 44. The time required for rulesets, however, increased from 32 to 9,715 s, a factor of 300.

3.2 IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ----> (Cause—of--- disease)
Example 1: If_then_rule induced in the diagnosis of level of alcohol in blood

IF Sex = MALE AND Unit = 8.9 AND Meal = FULL THEN

Diagnosis=Blood_alcohol_content_HIGH.

3.3 Transformation parameters

For the use of data mining techniques in health care, we must transform the data according to the requirement. Heart attack related data records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database In order to set the transformation parameters we must discuss attributes corresponding to heart vessels. The LAD, RCA, LCX and LM numbers represent the percentage of vessel narrowing (or blockage) compared to a healthy artery. Attributes LAD, LCX and RCA were partitioned by cutoff points at 50 and 70%. In the cardiology field, a 70% value or higher indicates significant coronary disease and a 50% value indicates borderline disease. A value lower than 50% means the patient is healthy. The most common cutoff value used by the cardiology community to distinguish healthy from sick patients is 50%. The LM artery is treated different because it poses higher risk than the other three arteries. Attribute LM was partitioned at 30 and 50%. The reason behind these numbers is both the LAD and the LCX arteries branch from the LM artery and then a defect in LM is more likely to cause a larger diseased heart region. That is, narrowing (blockage) in the LM artery is likely to produce more disease than blockages on the other arteries. That is why its cutoff values are set 20% lower than the other vessels. The nine heart regions (AL, IL, IS, AS, SI, SA, LI, LA, AP) were partitioned into two ranges at a cutoff point of 0.2, meaning a perfusion measurement greater or equal than 0.2 indicated a severe defect. Cholesterol CHOL was partitioned with cutoff points 200 (warning) and 250 (high). These values correspond to known medical settings. Decision tree rules with numeric dimensions and automatic splits.

Predicting diseased arteries:

IF ($SA \leq 0.37$ $AP > 0.66$)
THEN $LAD \geq 50$ ls=10% cf=0.80
IF ($SA \leq 0.37$ $AP <= 0.66$ $Age > 78$)
THEN $LAD \geq 50$ ls=4% cf=0.74
IF ($SA > 0.37$ $Age \leq 53$ $AS \leq 0.67$)
THEN $LAD \geq 50$ ls=1% cf=0.85
IF ($SA > 0.37$ $Age > 53$)
THEN $LAD \geq 50$ ls=8% cf=0.98

3.4 Neural Network Architecture

The architecture of the neural network used in this study is the multilayered feed-forward network architecture with 20 input nodes, 10 hidden nodes, and 10 output nodes. The number of input nodes is determined by the finalized data; the number of hidden nodes is determined through trial and error and the number of output nodes is represented as a range showing the disease classification. In general, results of disease classification or prediction task are true only with a certain probability.

3.4.1. Neuro-Fuzzy

Stochastic back propagation algorithm is used for the construction of fuzzy based neural network. The steps involved in the algorithm are as follows: First, initialize weights of the connections with random values. Second for each unit compute net input value, output value and error rate. Third, to handle uncertainty for each node, certainty measure (c) for each node is calculated. Based on the certainty measure the decision is made. The level of the certainty is computed using the following conditions.

If $0.8 < c \leq 1$, then there exists very high certainty
If $0.6 < c \leq 0.8$, then there exists high certainty
If $0.4 < c \leq 0.6$, then there exists average certainty
If $0.1 < c \leq 0.4$, then there exists less certainty
If $c \leq 0.1$, then there exists very less certainty

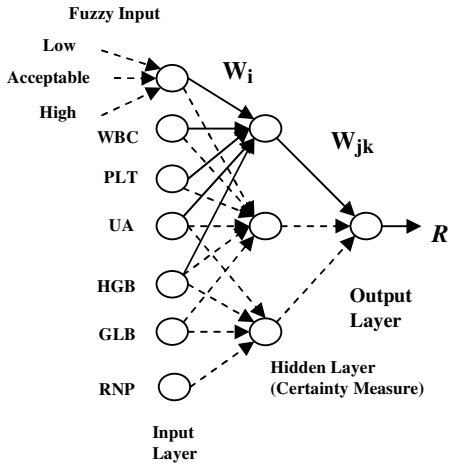

Figure 6. Trained Neural Network for Thrombosis

The network constructed consists of 3 layers namely an input layer, a hidden layer and an output layer. Sample trained neural network consisting of 7 input nodes, 3 hidden nodes and 1 output node is shown in Figure 6. When a thrombus (***R***) or blood clot occupies more than 75% of surface area of the lumen of an artery then the expected result may be a prediction of cell death or heart disease according to medical guidelines i.e**. *R*** is generated with reference to the given set of input data.

3.5. Bayesian Network Structure Discoveries

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*. This probability is denoted as P*(C | E)* where

$$P(C|E) = \frac{P(E|C).P(C)}{P(E)}$$

Bayes' theorem is the method of finding the converse probability of the conditional,

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

This conditional relationship allows an investigator to gain probability information about either C or E with the known outcome of the other. Now consider a complex problem with n binary variables, where the relationships among them are not clear for predicting a single class output variable (e.g., node 1 in Figure 7). If all variables were related using a single joint distribution, the equivalent of all nodes being first level parents, the number of possible combinations of variables would be equal to (2^n-1). This results in the need for a very large amount of data [Yoshinori Yaginuma]. If dependence relationships between these variables could be

determined resulting in independent variables being removed, fewer nodes would be adjacent to the node of interest. This parent-node removal leads to a significant reduction in the number of variable combinations, thereby reducing the amount of needed data. Furthermore, variables that are directly conditional, not to the node of interest but to the parents of the node of interest (as nodes 4 and 5 are with respect to node 1 in Figure 7), can be related, which allows for a more robust system when dealing with missing data points. This property of requiring less information based on pre-existing understanding of the system's variable dependencies is a major benefit of Bayesian Networks [Neapolitan, R.]. A Bayesian Network (BN) is a relatively new tool that identifies probabilistic correlations in order to make predictions or assessments of class membership.

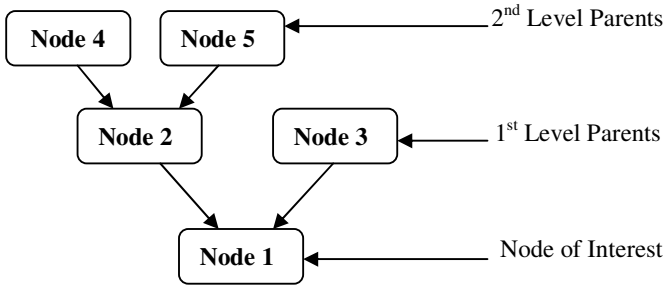

Figure 7. Decision tree rules with numeric dimensions and automatic splits

3.5.1 Basic Bayesian Network Structure and Terminology

While the independence assumption may seem as a simplifying one and would therefore lead to less accurate classification, this has not been true in many applications. For instance, several datasets are classified in [Domingos, P. and M. Pazzani] using the naïve Bayesian classifier, decision tree induction, instance-based learning, and rule induction. These methods are compared showing the naïve classifier as the overall best method. To use a Bayesian Network as a classifier, first, one must assume that data correlation is equivalent to statistical dependence.

3.5.2. Bayesian Network Type

The kind of Bayesian Network (BN) retrieved by the algorithm is also called Augmented Naïve BN, characterized mainly by the points below

- All attributes have certain influence on the class
- The conditional dependency assumption is relaxed (certain attributes have been added a parent)

3.6. Analysis of Results

In our previous work these models are evaluated by obtaining of 897 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database foundation. Every model was evaluated based on the two measures discussed above (classification accuracy). The results were achieved using average value of tenfold cross-validation for each algorithm. As shown in Figure 8, Bayesian model (BN) achieved classification accuracy of 0.82 with a sensitivity of 0.87. The SVM achieved classification accuracy of 0.835 with a sensitivity of 0.88 as shown in Table-3.

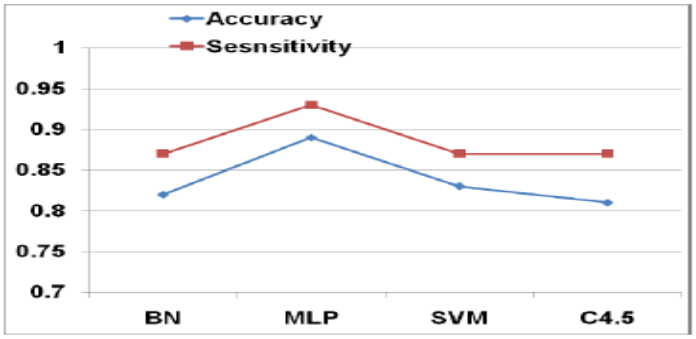

Figure 8. Comparison of four models, Neural Network (MLP) proving highest accuracy

The decision trees (C4.5) achieved a classification accuracy of 0.82 with a sensitivity of 0.88 for the data not having disease. However, the neural network model (MLP) preformed the best of the four models evaluated. MLP achieved a classification accuracy of 0.8920 with a sensitivity of 0.92.

Table 3. Analysis of data with heart disease.

Algorithm	**Attributes used**	**Sensitivity %**	**Accuracy %**
Decision Tree(C4.5)	Without Disease	88	82
Neural Network(MLP)	Having Disease	**92**	**89.2**
Bayesian model	Having Disease	87	82
SVM	Having Disease	88	83.5

Why neural network (MLP) performs best? We know that syndrome is combination of symptoms. Indeed, syndrome is a diagnosed concept produced by mean of mapping symptoms to TCM expert's brains. So syndrome is identified by human brain and neural network is considered as best modeler of human brain. Furthermore, neural network can approximate arbitrary mapping, while syndrome is a mapping of symptoms. These two reasons may explain why neural network is with best performance.

4. TIME SERIES ANALYSIS

The first step in traditional outbreak detection methods is to develop a model that can describe the normal time series patterns. The most widely used model is the Autoregressive Integrated Moving Average (ARIMA) models of Box and Jenkins. The model setting can be described by three parameters: (*p, d, q*). The parameter *p* refers to the length of historical time series values that can affect current observations. The second parameter *d* specifies how many different operations are required to make the time series stationary. The third parameter *q* specifies the length of historical error terms that can affect current observations. In a typical setting that does not involve seasonal fluctuation, the observed time series is usually assumed to be stationary, that is, $d = 0$. Specifically, an ARIMA (p, 0, q) model can be written as:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \epsilon_t + b_1 \epsilon_{t-1} + \ldots + b_q \epsilon_{t-q},$$

where y_t is the observed time series and ϵ_t is the error term. To ensure that the model "learns" the normal time series pattern, the data used for model estimation should be outbreak-free. Given p and q, the parameter values $(a_0, a_1, \ldots b_q)$ can be estimated using likelihood maximization. However, different model settings that correspond to different values of p and q may affect prediction accuracy. The values of p and q are usually determined by model selection criteria that take both goodness of fit and model complexity into consideration. Commonly used model selection criteria include Akaike information criterion (AIC) and Bayesian information criterion(BIC). Note that the model selection criteria are closely related to the "cross-validation" evaluation approach commonly use by the machine learning community. In fact, cross-validation is asymptotically equivalent to AIC. Other modeling techniques such as the generalized linear model using Poisson distribution, expectation-variance model, and the Wavelet model have been evaluated in previous studies. For the purpose of detecting outbreaks, there are two issues warranting further discussion the modeling of the day-of-week and seasonal effects.

4.1. ARIMA Time Series Analysis

The Box-Jenkins or Autoregressive Integrated Moving Average (ARIMA) methodology involves finding solutions to the difference equation

$$\phi_p(B)\phi_p(B^L)x_t = \delta + \theta_q(B)\theta_Q(B^L)a_t$$

- The nonseasonal autoregressive operator $\phi_p(B)$ of order p models low-order feedback responses.
- The seasonal autoregressive operator $\phi_p B^L$ of order P models feedback responses that occur periodically at seasonal intervals. For example, given a time series of monthly data, this operator would be used to model a regressive effect that occurs every January.
- The nonseasonal moving average operator $\theta_q(B)$ of order q models low-order weighted average responses.
- The seasonal moving average operator $\theta_Q B^L$ of order Q models seasonal weighted average responses.
- The terms x_t , a_t, and δ are the time series, a sequence of random shocks, and a constant, respectively.

The orders of the operator are selected ad hoc, and the parameters are calculated from the time series data using optimization methods such as maximum likelihood and least squares. The ARIMA method is limited by the requirement of stationary and invertibility of the time series i.e., the system generating the time series must be time invariant and stable. Additionally, the residuals, the differences between the time series and the ARIMA model, must be independent and distributed normally. Although integrative (filtering) techniques can be useful for converting nonstationary time series into stationary ones, it is not always possible to meet all of the requirements.

The ARIMA model is best presented in terms of the following operators. The backshift operator B shifts the index of a time series observation backwards, e.g. $BZ_t = Z_{t-1}$ and $B^k Z_t = Z_{t-k}$. The nonseasonal or first difference operator, $\nabla = 1 - B$, provides a compact way of describing the first difference. The seasonal operator ∇_L is useful for taking the difference between two periodic or seasonal time series observations. It is defined as $\nabla_L = 1 - B^L$

Having introduced the basic operator notation, the more complex operators presented and they can be discussed. The first operator form is the nonseasonal autoregressive operator $\phi_p(B)$,

also called the "Green's function". This operator captures the systems dynamical response to a_t – the sequence of random shocks – and previous values of the time series. The second operator is the nonseasonal moving average operator $\theta_q(B)$. It is a weighted moving average of the random shocks a_t.

The third operator is the seasonal autoregressive operator $\emptyset_P(B^L)$. It is used to model seasonal regressive effects. For example, if the time series represents the monthly sales in a toy store, it is not hard to imagine a large increase in sales just before Christmas. This seasonal autoregressive operator is used to model these seasonal effects. The fourth operator is the seasonal moving average operator $\theta_Q(B^L)$. It also is useful in modeling seasonal effects, but instead of regressive effects, it provides a weighted average of the seasonal random shocks. The constant $\delta = \mu\emptyset_p(B)\emptyset_P(B)$, where μ is the mean of the modeled stationary time series.

Bowerman suggests three steps to determine the ARIMA model for a particular time series.
1. Should the constant δ should be included?
2. Which of the operators $\emptyset_p(B)$, $\emptyset_P(B^L)$, $\emptyset_q(B)$, and $\emptyset_Q(B^L)$ are needed?
3. What order should each selected operator have?

The δ should be included if

$$\frac{\mu(Z)\sqrt{c(Z)}}{\sigma_X} > 2$$

Diagnostic checking of the overall ARIMA model is done by examining the residuals. The first diagnostic check is to calculate the Ljung-Box statistic. Typically, the model is rejected when the α corresponding to the Ljung-Box statistic is less than 0.05. For non-rejected models, the residual sample autocorrelation function (RSAC) and residual sample partial autocorrelation function (RSPAC) should have absolute *t* statistic values greater than two. For rejected models, the RSAC and RSPAC can be used to suggest appropriate changes to enhance the adequacy of the models.

4.2. Approach to Time Series Data Mining (TSDM)

The first step in applying the TSDM method is to define the TSDM goal, which is specific to each application, but may be stated generally as follows. Given an observed time series

$$X = \{x_t, t = 1, \ldots, N\}$$

the goal is to find hidden temporal patterns that are characteristic of events in *X*, where events are specified in the context of the TSDM goal. Likewise, given a testing time series

$$Y = \{x_t, t = R, \ldots, S\}\ N < R < S$$

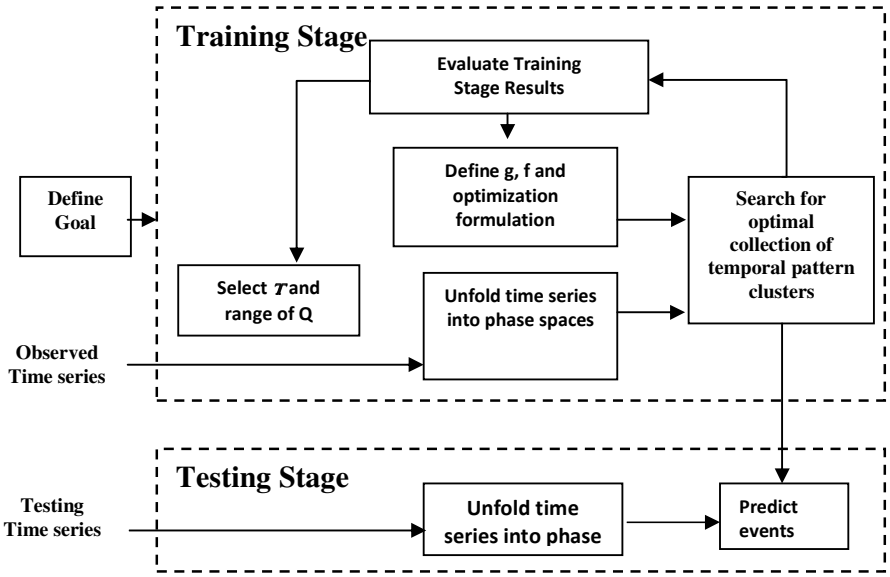

Figure 9. Block diagram for TSDM Method

the goal is to use the hidden temporal patterns discovered in *X* to predict events in *Y*.

Figure 9 presents a block diagram of the TSDM method. Given a TSDM goal, an observed time series to be characterized, and a testing time series to be predicted, the steps in the TSDM method are.

I. Training Stage (Batch Process)

1. Frame the TSDM goal in terms of the event characterization function, objective function, and optimization formulation.
   a) Define the event characterization function *g*.
   b) Define the objective function f.
   c) Define the optimization formulation, including the independent variables over which the value of the objective function will be optimized and the constraints on the objective function.
2. Determine Q, i.e., the dimension of the phase space and the length of the temporal pattern.
3. Transform the observed time series into the phase space using the time delayed embedding process.
4. Associate with each time index in the phase space an eventness represented by the event characterization function. Form the augmented phase space.
5. In the augmented phase space, search for the optimal temporal pattern cluster, which best characterizes the events.
6. Evaluate training stage results. Repeat training stage as necessary.

4.2.1. TSDM Training Step 1 – Frame the TSDM Goal in Terms of TSDM Concepts

The first step in the TSDM method is to frame the data mining goal in terms of the event characterization, objective function, and optimization formulation. Since the goal is to characterize the morbidity, the event characterization function is $g(t) = x_{t+1}$ which allows prediction one time-step in the future.

Since the temporal patterns that characterize the events are to be statistically different from other temporal patterns, the objective function is

$$f(P) - \frac{\mu_M - \mu_{\tilde{M}}}{\sqrt{\frac{\sigma_M^2}{c(M)} + \frac{\sigma_{\tilde{M}}^2}{c(\tilde{M})}}}$$

which orders temporal pattern clusters according to their ability to statistically differentiate between events and non-events. The optimization formulation is to max $f(P)$ subject to min $b(P)$ such that minimizing $b(P)$ does not change the value of $f(P)$. This optimization formulation will identify the most statistically significant temporal pattern cluster with a moderate radius. The function b determines a moderate δ based on an electrical field with each phase space point having a unit charge. The function b measures the cumulative force applied on the surface of the temporal pattern cluster.

4.2.2. TSDM Training Step 2 – Determine Temporal Pattern Length

The length of the temporal pattern Q, which is also the dimension of the phase space, is chosen ad hoc. Takens' theorem in figure 10 proves that if $Q=2m+1$, where m is the original state space dimension, the reconstructed phase space is guaranteed to be topologically equivalent to the original state space, but Takens' Theorem provides no mechanism for determining m. Using the principle of parsimony, temporal patterns with small Q are examined first. For this example, $Q = 2$, which allows a graphical presentation of the phase space.

Theorem (Takens) : Let the state space M of a system be Q dimensional, $\varphi: M \to M$ be a map that describes the dynamics of the system, and $y: M \to R$ be a twice continuously differentiable function, which represents the observation of a single state variable. The map $\Phi_{(\varphi,y)} = M \to R^{(2Q+1)}$ defined by

$\Phi_{(\varphi,y)}(x) = (y(x), y(\varphi(x)), \ldots, y(\varphi^{2Q}(x)))$

is an embedding. An embedding is a homeomorphic mapping from one topological space to another where a homeomorphic map is continuous, bijective (one to-one and onto), and its inverse is continuous.

Figure 10. Taken's theorem

4.2.3. TSDM Training Step 3 – Create Phase Space

The time series X is embedded into the phase space using the time-delay embedding process where each pair of sequential points (x_{t-1}, x_t) in X generates a two dimensional phase space point. If the phase space were three-dimensional, every triplet of sequential points (x_{t-2}, x_{t-1}, x_t) could be selected to form the phase space. The Manhattan or l_1 *d*istance is chosen as the metric for this phase space.

4.2.4. TSDM Training Step 4 – Form Augmented Phase Space

The next step is to form the augmented phase space by extending the phase space with the $g(.)$ dimension as a stem-and-leaf plot. The vertical lines represent the dimension g associated with the pairs of (x_{t-1}, x_t). The next step will find an optimal cluster of leaves with high eventness.

4.2.5. TSDM Testing Step 1 – Create Phase Space

The testing *time* series Y, which is shown in the nonstationary, nonperiodic continuation of the observed time series. The time series Y is embedded into the phase space using the time-delay embedding process performed in the training stage.

4.2.6. TSDM Testing Step 2 – Predict Events

The last step in the TSDM method is to predict events by applying the discovered temporal pattern cluster to the testing phase space.

5. CONCLUSION

We focused on using different algorithms for predicting combinations of several target attributes. In this paper, we have presented effective heart attack prediction methods using data mining techniques. Firstly, we have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack based on the calculated significant weightage. The frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals are defined based on data exploration. All these models could answer complex queries in predicting heart attack. The incidence of heart attack is increasing every year in coal mining regions, especially in India, is a constant intimidation to the population and a recurring problem for the health authorities. Forecasting heart morbidity epidemic can help the authorities to take effective measures to handle any unexpected situation. Forecasting methods are useful in healthcare management. Accurate prediction of patient attendances will facilitate timely planning of staff deployment and allocation of resources within a department or a hospital. The hospital where the study was carried out is a regional hospital, with its catchment of patients geographically determined. Largely, idea of forecasting is for a basis of macro-planning for taking relevant steps which up to now is based on an average, aggregated incremental percentage annual growth.

REFERENCES

[1] Á. Méri and L. Körmöczi, Temporal Pattern Analysis – A New Algorithm For Detecting Patch Size In Plant Populations, TISCIA 38, 3-9

[2] A. A. Freitas and S. H. Lavington, Mining very large databases with parallel processing. Boston: Kluwer Academic Publishers, 1998.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39:1.38, 1977.

[4] A. Szymkowiak, P.A. Philipsen, J. Larsen, L.K. Hansen, E. Thieden and H.C. Wulf, Imputating missing values in diary records of sun-exposure study, in Proceedings of IEEE Workshop on Neural Networks for Signal Processing XI, ed itor: D. Miller, T. Adali, J. Larsen, M. Van Hulle, S. Douglas, pages: 489.498.

[5] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan, Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques The 5th International Conference on Computer Science & EducationHefei, China. August 24–27, 2010

[6] Abdurazzag A ABURAS and Nurul Fariza ZULKURNAIN, Investigation of the Time-Series Medical Data based on Wavelets and K-means Clustering, ARISER Vol. 3 No. 3 (2007) 112-122

[7] Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.

[8] Box, G.E.P. and Jenkins, G.M. (1970), Time Series A nalysis: Forecasting and Control, San Francisco: Holden-Da y.

[9] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. A new method for similarity indexing of market basket data. In Proc. 1999 ACM SIGMOD Int. Conf. on Management of data, pages 407-418, 1999.

[10] C. J. Deschenes and J. P. Noonan. A fuzzy kohonen network for the classification of transients using the wavelet transform for feature extraction. Information Sciences, (87):247– 266, 1995.

[11] Domingos, P. and M. Pazzani, On the optimality of the simple Bayesian classifier under zeroone loss. Machine Learning, 1997. 29(2-3): p. 103-30.

[12] E. Keogh, "A Fast and Robust Method for Pattern Matching in Time Series Databases," proceedings of 9th International Conference on Tools with Artificial Intelligence (TAI '97), 1997.

[13] F. Abramovich, T. Bailey, and T. Sapatinas.Wavelet analysis and its statistical applications. JRSSD, (48):1–30, 2000

[14] Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A

[15] Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

[16] Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163

[17] G. E. P. Box and G. M. Jenkins, Time series analysis: forecasting and control, Rev. ed. San Francisco: Holden-Day, 1976.

[18] Grabmeier and Andreas Rudolph. Techniques of cluster algorithms in data mining. Data Mining and Knowledge Discovery, 6(4):303. 360, October 2002.

[19] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining. Boston: Kluwer Academic Publishers, 1998

[20] Hand D., Mannila H. & Smith P. (2001) Principle of Data Mining. MIT Press

[21] Hendryx and Ahern, 2008, Chronic Illness Linked To Coal-Mining Pollution, Study, ScienceDaily (Mar. 27, 2008)

[22] Lawrence Saul and Fernando Pereira. Aggregate and mixed-order Markov models for statistical language processing. In Claire Cardie and Ralph Weischedel, editors, Proceedings of the Second Conference on Empirical

[23] Michael Hendryx , Keith J. Zullig, Higher coronary heart disease and heart attack morbidity in Appalachian coal mining regions, journal homepage: www.elsevier.com/ locate/ypmed, Preventive Medicine 49 (2009) 355–359

[24] Naus. J (1965) A scan statistic with variable window, Statistics in Medicine 15,845-850

[25] Neapolitan, R., Learning Bayesian Networks. 2004, London: Pearson Printice Hall.Krishnapuram, B., et al., A Bayesian approach to joint feature selection and classifier design.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 6(9): p. 1105-1111.

[26] Nobre, F. F. and Stroup, D. F. (1994), "A monitoring system to detect changes in public health surveillance data," Journal of Epidemiology, 23, 408–418.

[27] P. Cabena and International Business Machines Corporation., Discovering data mining : from concept to implementation. Upper Saddle River, New Jersey: Prentice Hall, 1998.

[28] P. Gray and H. J. Watson, Decision support in the data warehouse. Upper Saddle River, New Jersey: Prentice Hall, 1998.

[29] Pedro Domingos , Michael Pazzani , On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, 29, 103–130 (1997) c° 1997 Kluwer Academic Publishers. Manufactured in The Netherlands pp.37-43.

[30] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In Proc. 4th Int. Conf. of Data Organization and Algorithms, pages 69-84. Springer, 1993.

[31] R. J. Povinelli and X. Feng, "Data Mining of Multiple Nonstationary Time Series," proceedings of Artificial Neural Networks in Engineering, St. Louis, Missouri, 1999, pp. 511-516.

[32] Rumelhart, D.E., McClelland, J.L., and the PDF Research Group (1986), Parallel Distributed Processing, MA: MIT Press, Cambridge. 1994.

[33] S. M. Pandit and S.-M. Wu, Time series and system analysis, with applications. New York: Wiley, 1983

[34] Steven J. Leon. *Linear Algebra with Applications*. Prentice Hall, 6th edition, January 2002.

[35] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings* Of The 22nd Annual Acm Conference On Research And Development *In Information Retrieval*, pages 50.57, Berkeley, California, August 1999.

[36] Tiao, G.C. and Tsa y, R.S.(1983), \Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models", The Annals of Statistics 11, 856-871.

[37] Uno T, Asai T, Uchida Y, Arimura H (2004) An efficient algorithm for enumerating frequent closed patterns in transaction databases. In: Proc. of the 7th international conference on discovery science. LNAI vol 3245, Springe, Heidelberg, pp 16–30

[38] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence. All rights reserved. 0738-4602-1996

[39] R.Kavitha kumar, dr. RM.Chadrasekaran, Attribute correction-data cleaning using association rule and clustering methods, International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.2, march 2011

[40] Batini Carlo, Barone Daniele, Cabitza Federico and Grega Simone , A Data Quality Methodology for Heterogeneous Data ,IJDMS, February 2011, Volume 3, Number 1

[41] Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Authors Profile

K.Srinivas is currently working as an Associate Professor at Jyothishmathi Institute for Technology & Science ,Karimnagar, AndhraPradesh, India .He received his M.Tech(Software Engg) degree in 2008 from Kakatiya University,Warangal and B.Tech.(CSE) degree in 1990 from Mysore University , Mysore. At present he is pursuing Ph.D(CSE) in JNTUH, Hyderabad. His areas of interest are incremental patterns, prediction using Data Mining Techniques.