DALARNA
University College

DEGREE PROJECT
COMPUTER ENGINEERING

Clustering Lab Values
Working with Medical Data

By
Mahtab Davari

A thesis
Presented to the University of Dalarna
In fulfillment of the
Thesis requirement for the degree of
Master of Computer Engineering
Borlänge, Dalarna, Sweden, 2007

Programme International Master Science in Computer Engineering (Specialization Intelligent Systems)	Reg.Number E3348D	Extent 30 ECTS
Name of student Mahtab Davari	Year-Month-day 2007-03-19	
Advisor Pascal Rebreyend	Examiner Mark Dougherty	
Company/Department the Friedrich-Schiller univeristy of Jena (university hospital)-Germany	Supervisor at company/Department Franziska Oroszi	
Title Clustering Lab Values Working with Medical Data		
Keywords Pneumonia, lab values, Data Mining, Clustering ,clustering algorithms, EM, K-means, Frequency of lab values, outliers.		

Abstract

Data mining is a relatively new field of research that its objective is to acquire knowledge from large amounts of data. In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available [27]. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. A major objective of this thesis is to evaluate data mining tools in medical and health care applications to develop a tool that can help make rather accurate decisions.

In this thesis, the goal is finding a pattern among patients who got pneumonia by clustering of lab data values which have been recorded every day. By this pattern we can generalize it to the patients who did not have been diagnosed by this disease whose lab values shows the same trend as pneumonia patients does.

There are 10 tables which have been extracted from a big data base of a hospital in Jena for my work .In ICU (intensive care unit), COPRA system which is a patient management system has been used**.** All the tables and data stored in German Language database.

Acknowledgment

My deepest appreciation goes to my supervisor Diipl-Kffr.Franziska Oroszi for her guidance and patient and support and encouragement through my thesis and Dr. Pascal Rebreyend for all his supports and guidance and cooperation between Wirtschaftswissenschaftliche Fakultät of the Friedrich-Schiller-Universidad (university hospital in Germany) and Högskolan (Dalarna university) that he provided for me.
I would like to thank my thesis readers, and especially my sincere appreciation goes to Professor Mark Dougherty for his support and encouragement and his understanding in my special situation. I also would like to thank you Dr. Hasan Fleyeh for all supports and advices and help provided.

And lastly, again my heart goes to Franziska Oroszi, without her patience and love. I would have never finished the thesis.

Table of Contents

Chapter 1

Introduction

Data Mining is a process of achieving knowledge from databases or any data storage. Data mining presents new perspectives for data collection analysis. By data mining, trends and patterns in data will be identified.

Data mining processes have required combinations of techniques, from statistics, machine learning, database technology, pattern recognition, information retrieval and spatial data analysis… [30]

Data mining, also known as knowledge discovery in databases, to find previously unknown and useful information from data .Data mining algorithm cannot operate on raw data, data mining process may need to extract, format and convert the raw data before invoking the data mining algorithm. The extraction information can usually help decision making. [12]

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. The ability to use these data to extract useful and new information for quality healthcare is crucial. [13]

Medical informatics plays a very important role in the use of clinical data. In such Discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place.

Structured query languages (SQL) are well known language with very little freedom for manipulations and it is useful for finding information, as long as the user knows perfectly what he or she is searching for. Once the user provides the Query the processor will provide the user with the exact answer that is required for the solution.

This thesis started with inserting 10 tables into MYSQL Administration .The maximum record was 17,000000, which had their own problems like, lack of memory and software crashing. This made user to redo or change some configurations or parameters in My SQL, such as buffer size, increasing maximum size for temporary table (tmp_table_size) in my.ini file or even change it to new version.

For this thesis Excel and SPSS have been applied for analyses and storing data and modifications and programming (scripts, Visual basic coding), transformations lab values,visualization,etc .Some times since Excel supports up to 65500, work had to be switched from Excel to SPSS, because by joining tables, a bigger table will be achieved, to 200000 records some times.Since the data was extracted directly from the original data base in hospital, so we had so many extra and unwanted data ,and so many more than 60% missing vales, which increased by deleting some unwanted values.

.In our research we need to focus in clustering algorithms to evaluate data. This thesis is dealing with patients who got one sort of disease which is called Pneumonia which is an inflammation of the lung caused by infection with bacteria, viruses, and other organisms The goal is to find a structure or pattern among pneumonia that can be used as symptoms for diagnosis by working on lab test data.

In each chapter from chapter 4 to 6 is trying to investigating to answer these questions:

1- After clustering any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why?

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

1.1 Motivation:

There are numerous data mining tools and methods available today. Although machine intelligence tools have been used for flying airplanes, sending rockets to space, the use of machine intelligence with health related databases has been limited. Machine intelligence can be used as a second opinion for clinical classification or clustering.

One of the tools has a built in preprocessing tool. A preprocessing tool is used to convert raw data into a format understandable by the data-mining algorithms. [27]

For this case pneumonia diagnosed by different methods, one method which we will focus on is the laboratory results which are prescribed by doctor .In this thesis, for start, some initial information got from data base, like age range, sex, measure the frequency of lab values for each lab data.

Several different methods have been applied ,because data had another problem ,like there was some times more than one value for lab data ,(meaning for one day some times you could find several different values) or some times no value for certain day. First start to replace these values with maximum, because the analyses of patients lab values show that maximum (out of range has more frequency) .Then, no promising results found for none of lab data. So, these values replaced with minimum value, no promising result achieved. They were replaced by minimum, maximum, median, but unfortunately, no promising clusters achieved.

Finally, it has been decided to replace values with their mean values and work on real values and their differences (Δ) and see which one is giving better results. Another problem with data was the high amount of their missing values which always makes the results accuracy decreased. (Some times more than 80%).

As soon as the method has been discovered, the preparations of data and analyses of data started. First of all, pneumonia patients has been clustered separately and non-pneumonia separately, and then by joining them, work with different samples was started to make decisions. [24]

Different clustering applications were used, SPSS, HCE, Weka, .By SPSS, no results has been achieved, because of lots of missing value.HCE for hierarchical clustering and Weka, for the rest of algorithms were used.[16]

1.2 Goals and Objectives:

The application of artificial intelligence in healthcare is rather new. The goal of this thesis is to show that data mining can be applied to the medical databases, which will predict or cluster data with a reasonable accuracy. For a good prediction or clustering (pattern recognition) the preparation has to be done properly and then make some conclusion in each step. Finally the results has to be compared to see if there is any connection between what has been achieved and what was expected to be achieved.

A number of clustering algorithms has been used in this work to show the drawbacks and advantages of them for these specific data. At the end of each lab values (the ones which showed the promising cluster) a table for comparison of different algorithm has been drown.

Chapter 2

Machine learning

Machine Learning tries to study the computer algorithms that will be improved through experience [44]. In machine learning range from data mining programs
Which discover general rules in large data sets, to information filtering systems that automatically learn users' interests. Machine learning can be used to develop systems resulting in increased efficiency and effectiveness of the system.
Machine learning is also called concept learning. That is, computers can learn patterns within the data. Machine learning is called successful when it can correctly find all the instances that have the right patterns. Some times Machine cannot categorize correctly all the instances due to high variations in attributes present in the data like what we have seen in operation of large neural network which is impossible to analyze because of the huge number of variables involved ,so even a correct answer achieved it is impossible to understand what neural network think. (One reason that neural network was not a good option for this thesis data due to the lots of variables and attributes).

2.1. MINING PROCESS

The aim of the database system is to store of raw data. Definition of the data mining is the process of inferring knowledge from the database system. This discovery process to get the knowledge consists of sequence of the steps. In figure 1 is showed the mining process steps and right after the brief definition of data mining concepts and these steps explanation are presented.[14][12]

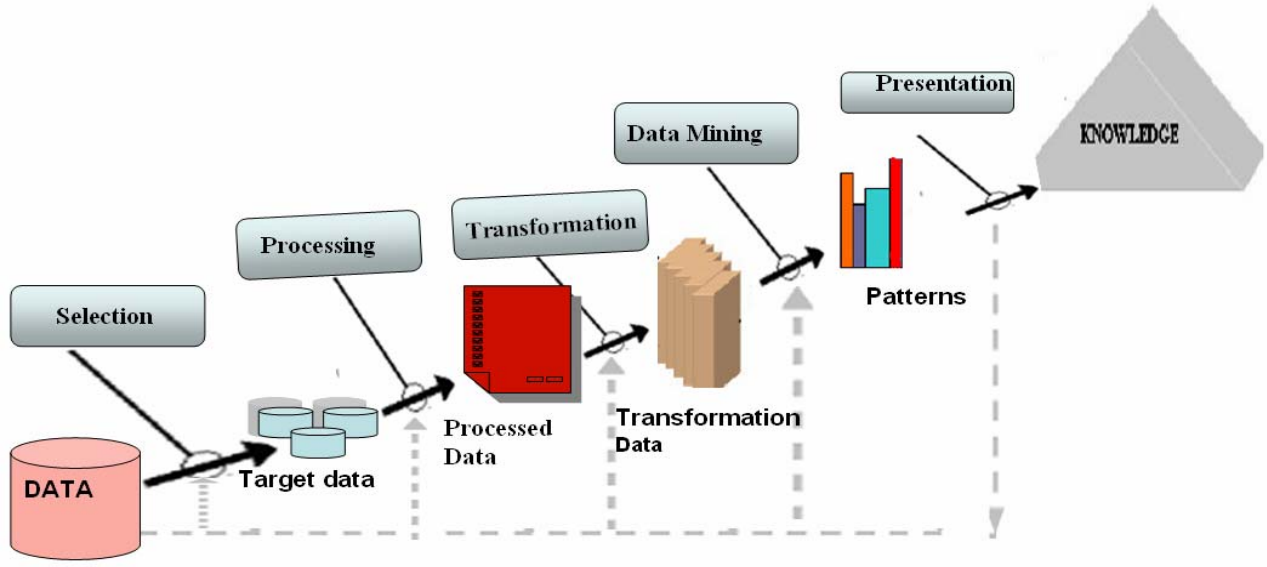

Figure 1. An Overview of the Steps That Composes the KDD Process.

So many tiring and complex mathematical or statistical calculations and finding specific information in a huge database can been done by using machines nowadays, by evolutions of machines we have possibility to use them for storing information, remind us of appointments, and so on. As the size of the data was increasing computer storage has increased. Since the vast amount of data that was being created, humans produced some algorithms that gives results a query is supplied. Although these tools work very well, but they can be applied to do only routine tasks. Automatic Classifications and clustering and other machine intelligence algorithms cannot be done using standard database languages. This has led to the creation of machine intelligence algorithms that can do tasks created by humans and make decisions without human supervision. From the evolution of machine intelligence came data mining. In data mining, algorithms seek out patterns and rules within the data from which sets of rules are derived. Algorithms can automatically

classify or clusters the data based on similarities (rules and patterns) .Data mining has grown that they can be used in many applications; such as predicting costs of corporate expense claims, in risk management, in process control in manufacturing, in healthcare, and in other fields.

2.1.1 Knowledge Discovery Process in databases [KDD] and data mining:

KDD is often used as a synonym for Data Mining; KDD is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data .(Fayyad, Piatetsky-Shapiro, and Smyth 1996). Knowledge discovery is the process of creating information automatically formalized in a way that is understandable for humans. Understanding and definition the problem are the first step in data mining process. Once a problem has been defined, relevant data must be collected. [31]
The relevant data is extracted from an existing database or data warehouse.

The steps involved in the KDD process (Figure.1):

Data extraction: A database is where all the data and information about the system is located. The step is to predefine our mission or a goal before discovering knowledge. Data for modeling can be stored in many different sources. These sources are databases, data-warehouses, web sites, text files; etc.A process that extracts useful subsets of data for mining is data extraction. Goals of the extraction processes are, identifying related information in the database and processing the database in a way which makes them properly ready to be analyzed by the data mining algorithms. Relevant data are collected from these sources during the extraction process. We have to select a subset of the database to perform the required knowledge discovery steps. Selection is the process of

selecting the right data from the database on which the tools in data mining can be used to extract information, knowledge and pattern from the provided raw data. [31]

Data Preprocessing: Large database systems usually contain errors in the stored data. Check the data for errors, outlines and missing values to the quality of the data. This is the most **time consuming and most important step** in the data preparation process. Robustness is an important property for the data mining systems. So, some techniques are used to manage to data in this process. (Small changes or some replacement (missing values) does not effect or change the result).Data cleaning is used to eliminate noise or error in the data.

Transformation: Data are transformed or integrated into the form which is suitable For mining, so we can say the data transformation process includes smoothing, generalization, normalization and aggregation techniques. Data reduction operation used to decrease the data size by using one of the data aggregation, dimension reduction or data comparison methods (usually there are cases where there are a high number of attributes in the database for a particular case. With the reduction of dimensionality we increase the efficiency of the data-mining step with respect to the accuracy and time utilization).

Data reduction methods can be used to minimizing representation of the data, while reducing the loss of information content.

Data mining: this step is the major step in data KDD. This is when the cleaned and preprocessed data is sent into the intelligent algorithms for classification, clustering, and so on, (for example, prepared data may contain many attributes and we have to select a subset of the attributes for using in data mining process). We chose the algorithms that

are good for discovering patterns in the data. . Data mining algorithm takes data as input and produces output in the form of models or patterns. In this step an intelligent methods are applied in order to extract data patterns.

Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right algorithms is important at this point.

Pattern: Data mining system creates lots of patterns, or rules. Only some of these patterns which are generated from the data mining system are interested to any given problem. Patterns have to be easily understood, useful, and interesting. An interesting pattern represents knowledge. After determination of the knowledge function, some measured functions which are used to separate uninteresting patterns from knowledge, are used for the data mining process. Thus, data mining algorithms use some constraints and measures to make sure the mining is complete. The optimization problem is one of the important problems for the data mining.

Knowledge Presentation: a kind of interpretation. In this step the mined data is presented to the end user in a human-viewable format. Visualization and Knowledge representation techniques are used to present the mined knowledge to user. In the knowledge presentation step some possible actions formed from the successful application.

2.2. Artificial Intelligence (AI) and Data mining

2.2.1 COMMON TOPICS IN DATA MINING AND AI [5]

Machine learning in AI is the most relevant area to data mining, from the AI perspective.

Three Fundamental AI Techniques in Data Mining: [28]

AI is a broader area than machine learning. AI systems are knowledge processing systems. Knowledge representation, knowledge acquisition, and inference including search and control, are three fundamental techniques in AI.

Knowledge representation. Data mining seeks to discover interesting patterns from large volumes of data. These patterns can take various forms, such as association rules, classification rules, and decision trees, and clustering therefore, knowledge representation becomes an issue of interest in data mining.

Knowledge acquisition. The discovery process shares various algorithms and methods with machine learning for the same purpose of knowledge acquisition from data or learning from examples.

Knowledge inference. The patterns discovered from data need to be verified in various applications and so deduction of mining results is an essential technique in data mining applications.

Therefore, knowledge representation, knowledge acquisition and knowledge inference, the three fundamental techniques in AI are all relevant to data mining.

Key Methods Shared in AI and Data Mining:

AI research is concerned with the principles and design of rational agents and data mining systems can be good examples of such rational agents. Most AI research areas (such as reasoning, planning, natural language processing, game playing and robotics) have concentrated on the development of symbolic and heuristic methods to solve complex problems efficiently. These methods have also found extensive use in data mining.

Symbolic computation. Many data mining algorithms deal with symbolic values. As a matter of fact, since a large number of data mining algorithms were developed to primarily deal with symbolic values, discretization of continuous attributes has been a popular and important topic in data mining for many years, so that those algorithms can be extended to handle both symbolic and real-valued attributes.

Heuristic search. As in AI, many data mining problems are NP-hard, such as constructing the best decision tree from a given data set, **and clustering a given number of data objects into an optimal number of groups.** Therefore, heuristic search, divide and conquer, and knowledge acquisition from multiple sources have been common techniques in both data mining and machine learning.

Knowledge discovery from large volumes of data is a research frontier for both data mining and AI, and has seen sustained research in recent years. From the analysis of their common topics, this sustained research also acts as a link between the two fields, thus offering a dual benefit. First, because data mining is finding wide application in many fields, AI research obviously stands to gain from this greater exposure. Second, AI techniques can further augment the ability of existing data mining systems to represent, acquire, and process various types of knowledge and patterns that can be integrated into many large, advanced applications, such as computational biology, Web mining, and fraud detection.

2.3 Types of Data Mining

Data mining consist of three major parts.

Clustering and classification: a set of data is analyzed and a set of grouping rules is created that can be used to classify future data.

Association: implies certain association relationships among a set of objects in a database.

Sequence analysis: seek to discover patterns that occur in sequence.

There are many different algorithms are used in these data mining for our case which is clustering. We will go through them.

2.3.1: Clustering goal

Clustering analysis is a sub-field in data mining that is used for finding similar groups in a large database. It has been applied in many fields of study, from ancient Greek astronomy to present-day insurance industry and medical. This thesis is dealing with a data base which has been recorded in a hospital .The data which are dealing with are the values of some test which has been done for each patients. These data are supposed to be day to day recorded, but for some reasons (errors which always happened) are some times more than 3 or more .These data some times has been stored in string format ,some times in numerical. We also can see some other values which are the non-relevant data (when the material has been finished, drop has been taken…).

In this thesis, we propose some different algorithms – the aim is finding a promising pattern among pneumonia patients to use it as symptoms for future work. These patterns might be found among non-pneumonia (who has not been diagnosed with this disease), we can easily decide since they have these symptoms, therefore they were vulnerable.

2.3.2 Clustering

We always dealing with two concepts, classifications and clustering.Clusterig are the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait –they have hidden information or pattern. Classification is a kind of categorization: the act of distributing things into classes or categories of the same type .

Cluster analysis (unsupervised learning), in which the classes are unknown a priori

The goal is to discover these classes from the data, but Classification (class prediction, supervised learning), the classes are predefined, the goal is to understand the basis for the Classification from a set of labeled objects and build a predictor for future unlabeled observations

The goal of the clustering is to take a set of objects which are records in a database and to partition them into number of groups or clusters.[20] (Berkhin P. & Software A., 2002). How to decide to use clustering pattern is the task for domain expert (*domain experts* display information in a logical fashion so that it can be easily coded into a computer system, allowing ease of use by end-users of the KBS systems) and data mining analyst. Cluster analysis method has been used in pattern recognition, data analysis and image processing. After clustering, we can apply some methods to discover rules predicting in a given class (when cluster is not unsupervised learning)

A good cluster is a group of data that has high 'Intra-cluster' similarity, but low –Inter-cluster' similarity. In other words, the members in a cluster own more similarities among themselves than to objects that belong to another group. The similarity can be physical or abstract. For example, a cluster could be the person who has specific disease, it could also

be a collection of images, etc. that share the same designation to form an abstract cluster. Clustering analysis unlike classification analysis does not need pre-defined group information. In other words, clustering analysis is “learning from observation instead of "learning from examples". This is the reason why clustering is referred to as "Unsupervised Learning" in artificial intelligence. But as soon as the pattern has been found and clusters achieved in visualization data and generalization and analysis of cluster, learning from data is not unsupervised it is supervised, For example, when we can label clusters achieved as a pneumonia patients (having pattern, promising result) and non-pneumonia cluster.

Clustering analysis is particularly useful when we have no labels for the classes, when we are interested in the inherited properties of the data, or when we know very little about the process creating the clusters.

Clustering can be divided into two basic types: hierarchical and partitional clustering. Within each of the types there exists a wealth of subtypes and different algorithms for finding the clusters.

Hierarchical clustering is either merging smaller clusters into larger ones, or splitting larger clusters. The clustering methods differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained.

Partitional clustering, on the other hand, attempts to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters.

2.3.3 Clustering Methods

The hierarchical algorithms: result is shown in a tree-like dendrogram. At the top of the tree each observation is represented as a separated “cluster” and at intermediate levels observations are grouped into fewer “cluster” than at the higher levels at the bottom, all of the observations are merged into one” cluster”.

The hierarchical clustering algorithm forms clusters in a hierarchical fashion. That is, the number of clusters at each stage is one less than the previous one. If there are n observations then at Step 1, Step2, …, Step n-1 of the hierarchical process the hierarchical process the number of clusters, respectively, will be n-1, n-2, …, 1.
Frequently the various steps or stages of the hierarchical clustering process are represented graphically in what is called a dendrogram or tree.
In fact, the various hierarchical clustering algorithms or methods differ mainly with respect to how the distances between the two clusters are computed. Some of methods which have been applied in this thesis are:

1. **Centroid method**: computes the distance between two clusters as the difference between centroids. The centroid of a cluster is the average point in the multidimensional space.

2. **Nearest-neighbor or single-linkage method**: In the single-linkage method, the distance between two clusters is represented by the minimum of the distance between all possible pairs of subjects in the two clusters.

3. **Farthest-neighbor or complete-linkage method:** The complete-linkage method is the exact opposite of the nearest-neighbor method. The distance between two clusters is defined as the maximum of the distances between all possible Pairs of observations in the two clusters

4. **Average-linkage method**: In the average-linkage method the distance between two clusters is obtained by taking the average distance between all pairs of subjects in the two clusters.

The non-hierarchical algorithms: I describe the one which has been used in this thesis:

EM: The Expectation-Maximization [2]

In Weka [16]: (EM) algorithm is part of the Weka clustering package. EM is a statistical model that makes use of the finite Gaussian mixtures model. Parameters are re-computed until a desired convergence value is achieved. The finite mixtures model assumes all attributes to be independent random variables. A mixture is a set of **N** probability distributions where each distribution represents a cluster. An individual instance is assigned a probability that it would have a certain set of attribute values given it was a member of a specific cluster. In the simplest case **N=2**, the probability distributes are assumed to be normal and data instances consist of a single real-valued attribute. Using

the scenario, the job of the algorithm is to determine the value of five parameters, specifically:

1. The mean and standard deviation for cluster 1
2. The mean and standard deviation for cluster 2
3. The sampling probability *P* for cluster 1 (the probability for cluster 2 is *1-P)*

Procedure is simplified as follows in general:

1. Guess initial values for the parameters. (μ_0, δ_0, *P,P-1*)
2. Use the probability density function for a normal distribution to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation δ, the formula is:

$$f(x) = \frac{1}{(\sqrt{2\pi})\, e^{\frac{-(x-\mu)^2}{2\delta^2}}}$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

3. Use the probability scores to re-estimate the five parameters.
4. Return to Step 2

The algorithm will be finished when a formula that measures cluster quality no longer shows significant increases. One measure of cluster quality is the likelihood that the data came from the dataset determined by the clustering. The algorithm finds the distribution parameters that maximize a model quality measure, called log likelihood.

The likelihood computation is simply the multiplication of the sum of the probabilities for each of the instances. With two clusters **A** and **B** containing instances

$x_1, x_2, \ldots x_n$ where $P_A = P_B = 0.5$

The computation is:

$$[0.5P(x_1|A)+0.5P(x_1|B)][0.5P(x_2|A)+0.5P(x_2|B)]...[0.5P(x_n|A)+0.5P(x_n|B)]$$

[2]

The reasons to choose EM clustering algorithm are: It has a strong statistical basis. It is linear in database size. It is robust to noisy data. It can accept the desired number of clusters as input. It provides a cluster membership probability per point. It can handle high dimensionality. It converges fast given a good initialization. [3] [25]

MakeDensityBasedClusterer: Is a wrapper for a cluster algorithm (EM, K-means, and Cobweb) used to return distribution and density information.
It fits normal distributions and discrete distributions within each cluster produced by the wrapped clustered. It supports the Number of Clusters requestable interface only if the wrapped

Cobweb generates hierarchical clustering, where clusters are described probabilistically. It works incrementally, updating the clustering instance by instance. The clustering COBWEB creates is expressed in the form of a tree, with leaves representing each instance in the tree, the root node representing the entire dataset, and branches representing all the clusters and sub-clusters within the tree. COBWEB starts with a tree consisting of just the root node. From there, instances are added one by one, with the tree

updated appropriately at each stage. When an instance is added, one of four possible actions is taken: The option with the greatest category utility is chosen. Category utility is defined by the function:

$$CU(C_1,C_2,\ldots C_k)=\frac{\sum_\ell Pr[C_\ell]\sum_i\sum_j(Pr[a_i=\nu_{ij}|C_\ell]^2-Pr[a_i=\nu_{ij}]^2)}{k}$$

$(C_1,C_2,\ldots C_k)$ the k clusters; the outer summation is over each of the clusters, which is later divided by k to provide a ``per cluster'' figure; the next inner summation sums over the attributes, and the inner-most summations sums over the possible values;

a_i is the ith attribute, and it takes on values $\nu_{i1},\nu_{i2,\ldots}$ Which are dealt with by the sum over j.

$Pr[A]$ Refers to the probability of event **A** occurring, $Pr[A|B]$ refers to the probability of event A occurring conditional on event B

Thus the difference $Pr[a_i=\nu_{ij}|C_\ell]^2-Pr[a_i=\nu_{ij}]^2$ refers to the difference between the probability that a_i has value ν_{ij} for an instance in cluster C_ℓ ,and the probability that a_i has value ν_{ij} .The larger this value, the more good the clustering does in terms of classification. This category utility formula only applies to categorical attributes (if the set $\{\nu_{i1},\nu_{i2,\ldots}\}$ would be infinite, and the summation could not be evaluated by conventional evaluation of a summation so:

$$\sum_i \sum_j \left(Pr[a_i = v_{ij} | C_\ell]^2 - Pr[a_i = v_{ij}]^2 \right) \Leftrightarrow \sum_i \left(\int f(a_i | C_\ell)^2 \, da_i - \int f(a_i)^2 \, da_i \right)$$

Where

$$\sum_i \left(\int f(a_i | C_\ell)^2 \, da_i - \int f(a_i)^2 \, da_i \right) = \frac{1}{2\sqrt{\pi \delta_{i\ell}}} \sum_i \left(\frac{1}{\delta_{i\ell}} - \frac{1}{\delta i} \right)$$

Sub-clustering and the tree structure can make COBWEB easier to understand than others. The order instances are read can greatly impact the clustering, sometimes placing two instances that are very similar and appear as the first input instances at opposite ends of the tree. [4].

K-mean: K-means is the unsupervised learning algorithms for clustering problem. First is defined k centroid, one for each cluster. Choosing centroids is better to place them as much as possible far away from each other. Then to take each point belonging to a given data set and associate it to the nearest centroid. When no point remained the early grouping is done. At this point we need to re-calculate k new centroids as centers of the clusters from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid (a loop). We may see that the k centroids change their location step by step until no more changes are done (centroids do not move any more).

Finally, this algorithm should minimize the *objective function*, (squared error function).

Which is $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$ where $\left\| x_i^{(j)} - c_j \right\|^2$ Chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centers. [6]

The Farthest First: Traversal Algorithm is a best possible heuristic for the k-center problem. To find k cluster centers, randomly choose one point as the first center

For L=2 upto k, next center = point with maximal min-distance to current centers.

In **Hochbaum-Shmoys** algorithm as shown below: Starting with an arbitrarily chosen point x_1 and adding it to the set *X*, the algorithm in each iteration picks that point in c_j which is farthest away from the points in *X* and adds it to *X*. At the end of *k* iterations, each point in *X* acts as a center for a stratum and the result of the algorithm is a disjoint partitioning obtained by assigning each point in c_j to its nearest point in *X*. the set of points in a patch c_j and a function $d(x, y)$ which gives the distance between any pair of points in that set. [7]

Data set $Y = \{y_1, y_2, \ldots y_n\}, X = \{\}$ **Hochbaum-Shmoys** algorithm

Pick an arbitrary point in Y and label it x_1

Add x_1 to X

For $i = 2,3,..k$

$$x_i = \arg\max_{y_p} \left[\min_q [d(y_p, x_q)] \right]$$

Add x_i to X.

For $J = 1,2,..k$

$$S_j = \left\{ y : y \in Y, x_j = \arg\min_{x_i} [d(y, x_\ell)] \right\}$$

Return X, $\{S_j\}$

Figure 2: The Hochbaum-Shmoys Algorithm. [7]

Chapter 3

3.1 Medical Data and Health informatics

Today, data mining has grown so vast that they can be used in many applications;
One important field is in healthcare. The number of people feeling sick and getting admitted into clinics and hospitals are increasing proportionally. The growing number of patients indirectly increases amount of data that are required to be stored. Some disease happened more than others in some part of hospitals, for example ICU (intensive care unit) which is the focus of this thesis. One of the goals of ICU is to diagnose the disease on time in order to cure them on time and also to decrease the day of staying patients in this center.

3.1.1 Medical DATA Descriptions:

In Order to have some view about the data which are supposed to be worked with, first of all we have to see what kind of tables we have and then join some tables for investigating information: Among 10 tables, 5 tables which have been chosen as follows:

Table Descriptions

1-"Faelle" (Cases) table: that consists of the patients personal identification like Patient ID (DatID),birthday, Sex ,name, clinical day acceptance, clinical day exit,..

2-"Laborwerte" (Lab value) table: that contains the date and time that lab has been prescribed by doctor and the date and time that is related to other tables

"labordefinitions" (lab definitions) and "laborgruppen" (lab test groups) also "labormengen" (lab quantities).

3-Laborgruppen (lab test groups) table: consists of the categories of lab data, including the name of the group and the index of the group of lab test.

4-Laborordefinition (lab definitions) table: consists of the definition of every single lab elements, their names and the scale for measuring, and Minimum and Maximum of Medical range of each lab elements.

5- Labormengen (lab quantities) table: consists of the results (quantity, value).

By joining them in MySQL and creating the new table.

Mining Process:

First Step: In order to work with lab data, we have to select the right sub set in selection part, First of all ,since there are main category (lab group),what has to be done is to arrange them in such a way that shows each group by its sub groups (any elements of lab test) .That was done by joining some above mentioned tables "labordefinitions" (lab definitions) and "laborgruppen" (lab test groups), which have the Group Name and Indexes Group in"laborgruppen" (lab test groups) and their each elements Name exists in "labordefinitions" (lab definitions) .At this point also the attributes that seems interesting for the samples is chosen ,by joining all above 5 tables .For the start:PatientID,Sex,age, Dates and time of lab value recorded, and lab values, and day of stay in ICU have been chosen, (in pneumonia case we have one extra attribute date that pneumonia happened).

Second and third Step: The goal in these steps is **Processing** (eliminating error) and does some **transformations** also by using the same table which mentioned in Step 1.

What has been noticed from the data extracted is, there are two fields in “labormengen” (lab quantities) that are having the lab value (quantity) achieved for each patients. One is called Wert (Value) which includes number (numeric Data type) and the other one is called VString, which includes text and numbers but has been stored as a string data type or combination of both.Vstring includes more errors and non relevant data such as, K.Mat, Mater, Proben (finished material, Probe is taken),….In order to make use of these two field, they have to be joined and to be unified to one data type .For clustering, Numeric data is our goal. Also the text like K.Mat and other errors should be removed. What else has been noticed that large amount of missing values (around 57% of lab data have been missed in only pneumonia patients in these two fields after joining).

All values are taken from the joined filed (Werte and Vsting) and are merged into the new assigned lab value by writing a scrip in Spss.When all new variables were filed it is the time to of cleaning of data should be started, which is the most important and time consuming part, needs to find all occurred cases. For example in some cases has inserted +, - in front or back of number, or used <,> or used comma for decimal part (+23, 12, 93---, <5.0, 2, 9).

Fourth and Fifth step: These steps are supposed to be a data mining step, pattern and knowledge representation, since we are separating lab values from each other and do cluster analysis and mining data(using different clustering algorithm).This step will be discussed for each lab value separately.

3.1.2. Pneumonia

This thesis is dealing with patients who got one sort of disease which is called Pneumonia. Pneumonia is an inflammation of the lung caused by infection with bacteria, viruses, and other organisms. Pneumonia is usually starts when a patient's defense system is weakened, most often by a simple viral upper respiratory tract infection or a case of influenza. People with infectious pneumonia often have a cough that produces greenish or yellow sputum and a high fever that may be accompanied by shaking chills and also Shortness of breath is common. To diagnose pneumonia, health care providers rely on a patient's symptoms and findings from physical examination. Information from a chest X-ray, blood tests, and sputum cultures may help.[1].Since pnuemonia is one of the three most important deadful infection in ICU(intensive care unit) and if it is diagnoised quickly it would be cured and patients can survive .Therefore ,it has a important role in ICU to deal with because by diagnoising on time and giving antibiotics (medications) the chance of healing will be increased and day of stay in ICU will be reduced.

The goal is to find a structure or pattern among pneumonia that can be used as symptoms for diagnosis by working on lab test data.

3.1.3. Initial Knowledge Extracted from Medical Data and Medical advices

Extracted pneumonia

In this part, it will show first the number of pneumonia and non-pneumonia that have been extracted from data base.

Number of pneumonia patients	630
Number of Non-Pneumonia Patients	17460

Table 1: number of patients

The Pneumonia patients have been extracted by using diagnosis table, a query that extracts patients which their ICD codes (The International Statistical Classification of Diseases and Related Health provides codes to classify diseases) are started with J18. [15].

Pneumonia and medical advice on lab data

Since the goal and the focus are to find a structure or pattern among pneumonia that can be used as the symptoms for diagnosis by working on lab test data. According to medical advices on lab data, which are interesting for physicians to be analyzed for pneumonia, are highlighted in table below (table 2).Table 2 is showing the total number of pneumonia patients that each lab values has been done for them, and as seen the table contains groups of lab data and their sub groups .According to medical decision, only focus on the patients who stayed in ICU more than 3 days.

Number of pneumonia patients		
group name	name	Total
AUTO Fibrinogen (a	Fibrinogen (abge	234
Blutgas arteriell	B.E.	518
(Blood gas arterial)	Ca++	518
	Cl-	518
	COHb	518
	Gluc	518
	Hb	518
	HbO2	518
	HCO3	518
	K+	518
	Lac	518
	MetHb	518
	Na+	518
	pCO2	518
	pH	518
	pO2	518
	SBC	518
Blutgas venös	B.E.v	374
	Chlorid.v	374
	Glukose.v	374
	Hb.v	374
	HbO2.v	374
	Kalium.v	374
	Kalzium.v	374
	Laktat.v	374
	Natrium.v	374
	pCO2.v	374
	pH.v	374
	pO2.v	374
	SBC.v	374
Endokrinologie	ACTH	192
	Angiotensin-C.E.	192
	Cortisol	192
	FSH	192
	Insulin	192
	Parathormon	192
	ß-HCG	192
	T3	192
	T3-frei	192
	T4	192
	T4-frei	192
	TSH	192
Herzenzyme	CK	378
	CK-MB	378
	CK-MB-Masse	378
	Troponin I	378
Leber-Pankreas	ALAT(GPT)	331
	Ammoniak	331
	AP	331
	ASAT(GOT)	331
	Bilirubin (dir)	331
	CD Transferrin	331
	CHE	331
	Gamma-GT	331
	GLDH	331
	LAP	331
	Lipase	331
	MEG-X1	331
	MEG-X2	331
	MEG-X3	331
Profila	Bilirubin (ges)	535
	CRP	535
	Harnstoff	535
	Kreatinin	535
Profilb	aPTT	536
	Erythrozyten	536
	Hämatokrit	536
	Hämoglobin	536
	INR	481
	Leukozyten	536
	MCH	481
	MCHC	481
	MCV	481
	Quick	536
	Thrombozyten	536
	U-Kreatinin	536
Profilc	PCT	488
Spiegel1	Amikacin	62
	Barbiturate	62
	Benzodiazepine	62
	Carbamazepin	62
	Cyclosporin	62
	Cyclosporin poly	62
	Digitoxin	62
	Digoxin	62
	Ethanol	62
	FK506	62
	Gentamycin	62
	Lithium	62
	MMF-Cellcept	62
	Phenobarbital	62
	Phenytoin	62
	Theophyllin	62
	Tobramycin	62
	Valproinsäure	61
	Vancomycin	61

Table2: List of lab data

Sex and age frequencies in pneumonia

Figure 3 shows age range in pneumonia and number of female and male in each range. The first frequent age range is 70-79 and the second most frequent is 60-69 and among them, male has the majority according to Figure 3.

We have to find out if sex or age can have an important role in clustering or not. (Later discussion in chapter 7 sex and age)

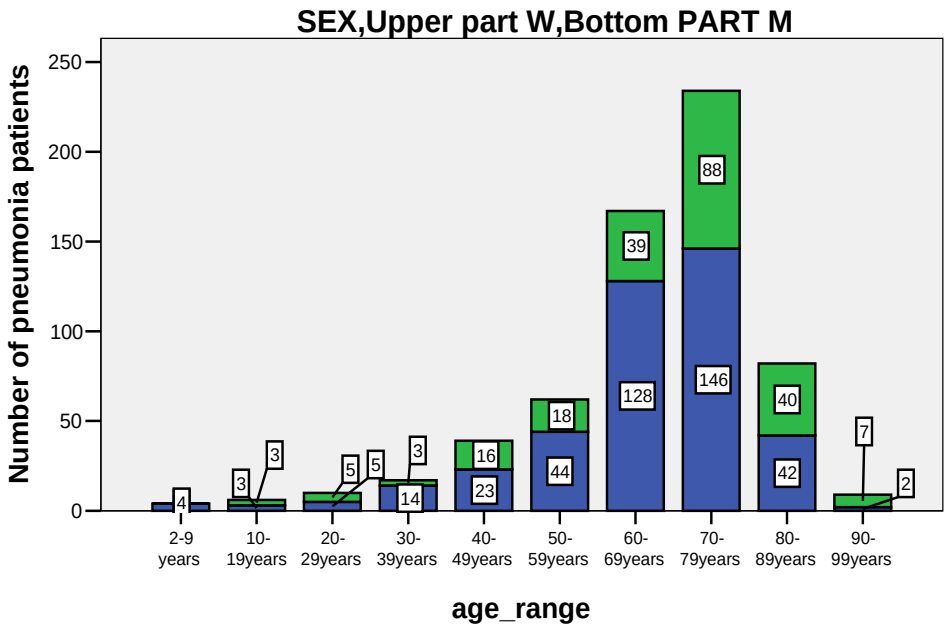

Figure 3: age and sex frequency in pneumonia patients

Chapter 4:

Clustering lab values

Each lab value analysis is started by getting to know the data, in terms of most frequent values in pneumonia and non-pneumonia and then start to do clustering separately (Data mining step) to learn how data are different in pneumonia and non-pneumonia (non-supervised learning) and find out the differences among clusters achieved. Right after ,by joining half pneumonia and non-pneumonia we will evaluate data (using Yes=pneumonia and No= not pneumonia)and use different Clustering algorithm .In this step, we will find out the best algorithm results (in terms of less incorrectly clustered and likelihood function and data distribution) and from that algorithm we find the promising cluster (more pneumonia and less non-pneumonia) .As soon as the promising cluster achieved we will look for difference between two clusters (the promising one and other cluster to see which attributes made a separations).

By looking back to the results from pneumonia separately, non-pneumonia separately and when we joining them, we try to find that the common critical day stay (common attributes that made the differences among clusters in pneumonia, non-pneumonia and joining of them).

Each chapter has frequency tables (showing the lab value range), the goal of this table and followed graph is to give us a view that how pneumonia and non-pneumonia for each lab data are close. So it will give us a confirmation to make decision of choosing the minor or major cluster as a promising result. In each lab value, two methods have been used.

One method is only work with lab value (real value) and another method works on differences (differences day stay k+1- day stay k).

What has been done to form the samples is to find the attribute that first both pneumonia and non pneumonias have in common .Second shows the real changes of lab values day to day specially the days close to when pneumonia started. The only attribute is the “day stay in ICU “but there are so many missing values in this data, first because that different patients stayed in ICU for different days and second because of lots of missing day lab values (according to what has been explained in previous chapter or the lab has not been not done in the specific day, or lots of mistakes in entering data or because some problem happened during test. The other thing that must consider is, according to what has been advised; only the patients who stayed in ICU more than 3 days should be taken into account.

If the good result (promising cluster) found, then that is tried to find any pattern in a smaller interval, for pneumonia patients the interval close to the day that the pneumonia has been diagnosed and some days after is more close to the interests. [11]

By looking at the data base and the data has been found out that several values for one day .Since we are supposed to work day to day so we have to replace them with a single value which can be one of the Max, Min, or median, Mean of these values. Based on the results that have been achieved that outrange of maximum is more common, so these values replaced with max value, but the clustering results were not reasonable, no promising results were achieved. By replacing these values by mean (average), better results were achieved. In the next part, only the lab data which gave us the promising results in clustering has been presented.

4.1 Blutgas arteriell .BE.

It is translated to the Arterial Blood Gas with sub component BE.(Basenexzess)Acid base balance.

4.1.1 Preliminary analyses of Blutgas.BE of Frequency lab value: [17] [12]

B.E_Pneumonia

Goal

In this part has been tried to find the most frequent lab values((most observed values)) in both pneumonia and non-pneumonia .Two tables achieved, one for pneumonia and for non-pneumonia in which their lab values range are presented by order of number of times each range has been observed. For example (2<=X<3, 164 times the lab values between 2 to 3 has been observed**,** and also number of patients that their lab values are in this range**,** see **Appendix B*: table 1 and table 2 Blutgas arteriell.B.E).***

Observations

By comparing these two tables we will see the most frequent lab values (most observed values) in two tables are the same and only some changes in frequency order are observed.

The goal is to find out how much pneumonia and non-pneumonia lab values are close .To see how lab values are distributed we just did some accumulation percentage over data. Figure 4 and Figure 5 are showing the percentage of distribution of lab values respectively in pneumonia and non-pneumonia. (Cumulative percentage of lab value distribution)

In both figures below, shows that more than 62 % of lab values in both pneumonia and non pneumonia are between -1 to 5.

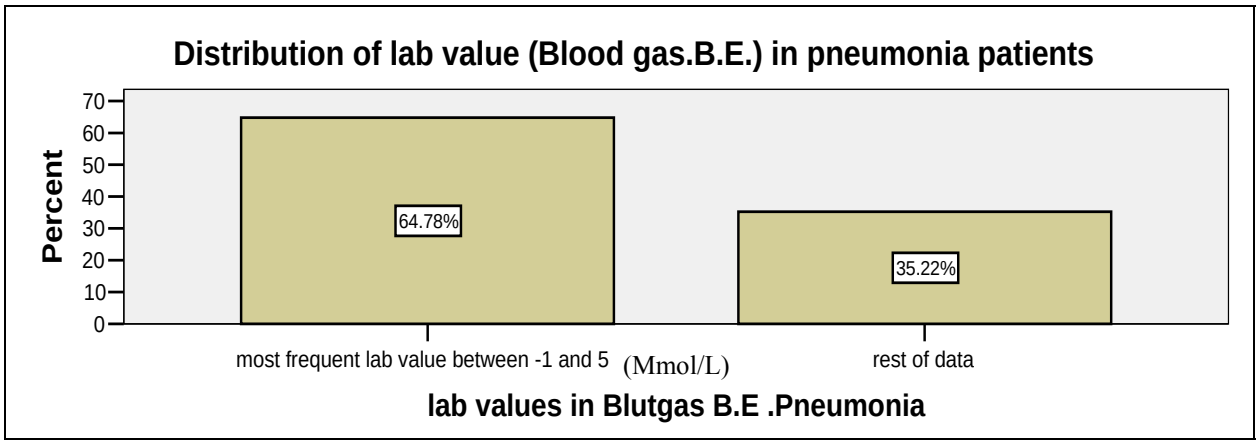

Figure 4: lab value distribution in Pneumonia of Blut gas.B.E.

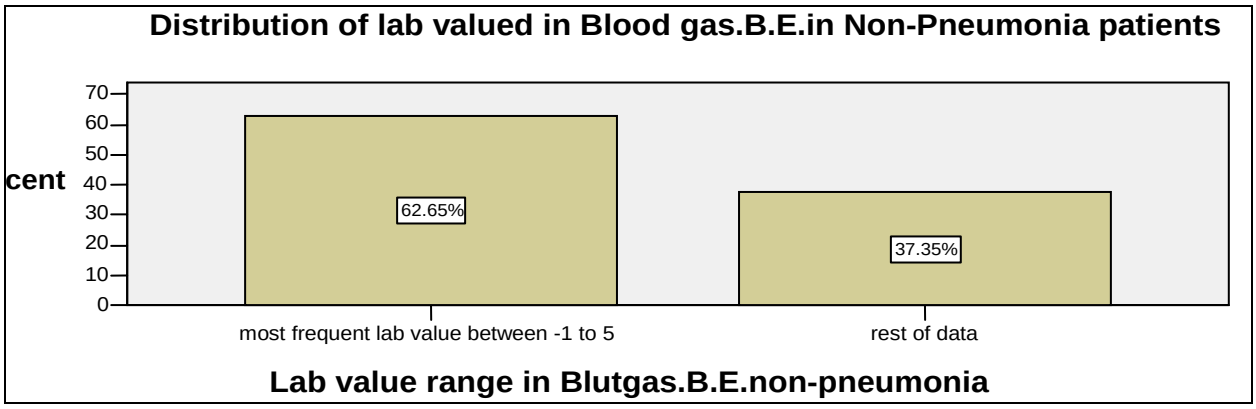

Figure 5: lab value distribution in Pneumonia of Blut gas.B.E.

Some hypothesis and assumption from 4.1.1.

1- We can expect after clustering (joining sample) more than 60% of data will form one cluster and the rest form another cluster. In other word, promising cluster has the less percentage of whole sample.

2- Since values are too close ,for finding promising and best cluster it may be better to apply differences (Δ) also (difference between two day stay e.g.daystay (k)-daystay (k+1)).

3- We can expect that, in clustering, probably only minor differences between promising cluster and other clusters will be observed.

4.1.2 Preliminary analyses of Blutgas.BE of out range of medical data:

The medical range that was inserted for this lab values are as **(-2.3 2.3) (**(mmol/L)**).**

Techniques

What has been done in this section is, the number of times that the Blood gas B.E has been tested for each patients have been sum up and then number of times that the value has been out of range of medical range was calculated and was found the percentage. (First column in two tables).

Results

In below tables as we see in table 3 and table4, the number of patients and their percentage of going out of range of Maximum of medical range is more probable, because the % number of patients that their labs values are out range of (2.3) is in the range 10 to 100% are 82% in pneumonia and 68% in non-pneumonia. Therefore, we can conclude that both pneumonia and no pneumonia cases have the tendency of going out range of maximum medical value (2.3), but in pneumonia, this tendency is more sensible, So we first decided to replace the multi value in one day with maximum, which after we found out we could not find any good result in clustering comparing with when we used average. These results are showing again how much lab values in pneumonia and non-pneumonia are close. (**For more information, see appendix B table 3, 4**).

Percentage of outrange (-2.3)min medical of lab value Blutgas.B.E	Number of patients in pneumonia (among 109)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 1313)	% number of patients in non-pneumonia
0-10%	82	75%	913	70%
10-100%	27	25%	400	30%

Table 3.Statistics value of Blutgas out range (-2.3) in range (-2.3 2.3). (Pneumonia and non-pneumonia patients)

Percentage of outrange 2.3 (max medical data) of lab value Blutgas.B.E	Number of patients in pneumonia (among 109)	% number of patients in pneumonia	Number of patients in non-pneumonia (1313)	% number of patients in non-pneumonia
0-10%	20	18%	437	32%
10-100%	80	82%	876	68%

Table4: Statistics value of Blutgas out range (2.3) in range (-2.3 2.3) (pneumonia and non-pneumonia patients)

4.1.3 Preparation stage of Blutgas.BE by separate clustering Pneumonia: [18]

a) Pneumonia (average, real value)

The first approach was started by clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) then next steps look at non-pneumonia and to see if any cluster will be found that helps for the final step (combination of pneumonia and non pneumonia) decisions ,e.g. is there any attributes that show a big difference among clusters? The Attributes are used for clustering are: 74 including (age, sex, 0daystay, 1 day stay… 71daystay).

The results achieved by using EM algorithm are as follows:

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	107	63.01583
1	19	

Table 5: Clustering results of pneumonia in Blutgas.B.E .using real values.

The results which shows the differences among clusters are in **appendix B table 5**.Since the results by using Δ ,showed the better and sensible result with bigger likelihood, just showed the results of them.

b) Pneumonia (Δ, difference of data average)

What we got from above tables in section a) made us to have a try on differences of day to day of lab value. Because it shows that there might be some changes from Day stay 15 … Day stay 23, which gives the better understanding or even better result. (The same 74 attributes are used: Age, sex, 0day, 1daystay …71daystay)

Thus, what can be done is to see how the changes are, so instead of studying lab value number itself, we calculated the differences (Δ) day to day and then try to cluster them again.

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	1	158.39471
1	13	
2	112	

Table 6: Clustering results of pneumonia in Blutgas.B.E .using differences (Δ) day to day.

Cluster differences

Table below shows the difference between clusters achieved in this clustering by EM.The interesting differences happened in Day stay 15 to Day stay 16 and Day stay 17 in cluster 1, We see jump from day stay 15 to 16 decreasing and then in Day stay 17 again jump to increasing while the changes are so minor in cluster 2.Cluster 0 is outlier it has only one observation.

Cluster numbers	Number of member	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD [19,21] Comparison of Cluster s by Δ			
		Day stay 15	**Day stay 16**	**Day stay 17**	**Day stay 18**
1	**13**	**0.5±0.3**	**-0.83±0.11** **As we see a big jump from day 15 to 16 (decreasing)**	**0.6±0.5** **(As we see a big jump from day 16 to day 17 increasing)**	**0.5±3**
2	**112**	**(0.45±0.6)**	**0.4±0.8** **Minor change**	**0.5±0.89**	**0.6±0.6** **Minor change(increase)**
0	**1**	**Not applicable (outlier)**			

Table 7: Summary of clustering all pneumonia of lab value Blutgas.B.E and its clusters differences using difference of mean value of data day to day Δ

3.1.4 Preparation stage of Blutgas.BE by separate clustering Non-Pneumonia:

a) Non-Pneumonia (real values)

The same way has been used here as we did in 3.1.3 a) but here For Non-pneumonia

Like in section a) the difference among clusters are better observed by using **Δ** and also the better log likelihood is achieved **,**the Δ results has been shown.(for observing the table of difference among clusters see **appendix B, table 6**).

The same attributes as before, 74 attributes: Age, sex, 0day, 1daystay, 2daystay…71daystay) and among 1062.

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	**97**	**-27.17162**
1	**115**	
2	**847**	
3	**3**	

Table 8: Clustering results of non-pneumonia in Blutgas.B.E .using **real value**.

b) Non- Pneumonia (Δ, difference of data average)

What can be done is to see how the changes are, so instead of studying lab value number, we calculated the differences (Δ) day to day of lab values and then try to cluster them again.

The same 74 attributes are used: Age, sex, 0day, 1daystay …71daystay)

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	**2**	39.0495
1	**121**	
2	**939**	

Table 9: Clustering results of non- pneumonia in Blutgas.B.E .using differences (Δ) day to day.

Difference of clusters

Table 10 shows the difference have been observed among obtained clusters .It is true that all the differences value are small but what made to highlight the cluster 3 is first of all, the cluster 3 in day stay 15 to day stay16, 17 comparing with other clusters can see some differences and in daystay16 shows a jump comparing with its similar cases in cluster 1 and 2 .So, cluster 3 it may be helpful in the final decision because in day stay 4 to day stay 6 and day stay 14 and day stay 16 and day stay 17 , the mean values show the significant difference with other clusters

Cluster	Number of member	**Comparison of Cluster s by Δ** **mean (daystayK+1 –Day stay K) ± STD** [19,21]						
1	121	**Day stay4 - Daystay3**	**Day stay5 –Daystay 4**	**Daystay6 – Daystay 5**	**Daystay14 – Day stay 13**	**Daystay15 – Daystay14**	**Daystay16 – Daystay 15**	**Day stay 17 – Daystay 16**
		0.06±1.8	**0.1±0.01**	**0.067±1.8**	**-0.01±0.17**	**0.03±1.6**	**0.01±1.5**	**0.08±1.5**
2	869	**0.14±1.18**	**-0.02±1.3**	**-0.04±1.1**	**-0.02±0.07**	**0.05±0.07**	**0.12±0.03**	**0.01±0.03**
3	70	**-0.24±1.3**	**0.307±1.18**	**0.6±0.3**	**0.3±0.12**	**0.4±1.6**	**- 0.7±0.2 (As we see big decreasing in day stay 16)**	**+0.5±0.03 Day stay 16----Day stay 17---- Day 18**
0	1	**Not applicable(outliers)**						

Table10: Summary of clustering all pneumonia of lab value Blutgas.B.E and its clusters Differences using difference of mean value of data day to day Δ

4.1.5 Clustering Joining Pneumonia and Non-pneumonia Blutgas.BE

Technique

In this step, the different samples have been tested (combination of 50% pneumonia, 50% non-pneumonia). Different algorithms have been applied and evaluated to find the best one, and also the data in samples has been tested in two aspects, lab values by itself and their differences. The preparation parts have been applied to make decision and according to the preparation results differences gave the better results in terms of what mentioned in Table 7 and also the pneumonia and non-pneumonia similar changes have be found in table 10, (so far common critical attributes are started from day stay 15 to day stay 16 and day stay 17).The same attributes have been used here 74: Age, sex, 0day, 1daystay...71daystay.

Results of technique

In Table 11 results have been sorted in terms of less incorrectly clustered and then bigger likelihood, and then number of clusters .The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

What has to be mentioned about Table ***11*** that it shows the difference between applying different algorithms. **The reason that the result from EM using differences has been chosen** are as follows:

1- The distribution of data in cluster 0(promising one) looks reasonable 33% of pneumonia and 8%non pneumonia (28, 6). (**A cluster with majority is with pneumonia**).
2- Comparing with other algorithms like EM(real value),K-means(real value and difference), Make density based cluster(real value and difference) ,which are having a promising cluster, **it has the better log likelihood function** (since likelihood measures how likely clustering is, so the greater the log likelihood is, the better the clustering result is).
3- **The percentage of incorrectly clusters instances are less than other algorithms**.

Algorithm	clusters	Pneumonia (84)	Non-pneumonia (84)	percentage	Data	**Log likelihood**	**Incorrectly clustered instances**	Number of clusters achieved
EM	0	28	6	20%	Differences	**167.377**	**35%**	2
	1	56	78	80%				
Make density based cluster	0	24	10	80%	Differences	**146.804**	**41%**	2
	1	60	74	20%				
EM	0	18	3	13%	real	**54.397**	**42 %**	3
	1	62	81	85%				
	2	4	0	2%				
Make density based cluster	0	10	3	8%	real	**40.30044**	**45 %**	2
	1	74	81	92%				
K-means	0	34	36	39%	real	-------	**45 %**	2
	1	50	48	61%				
K-means	0	29	36	39%	differences	----------	**45 %**	2
	1	55	48	61%				
Farthest first	0	83	84	99%	difference	----------	**46%**	2
	1	1	0	1%				
Farthest first	0	831	84	99%	real	----------	**49%**	2
	1		0	1%				
Hierarchical (Complete linkage ,Euclidean)	0	4	6	8%	Differences	----------	----------	2
	1	80	78	92%				
Hierarchical (Complete linkage ,Euclidean)	0	84	84	100%	real	**1**	---------	1
cobweb	non				Differences	------	**97%**	235
cobweb	non				real		**97%**	256

Table11: difference of applying different algorithm for clustering using Weka and HCE. Blutgas B.E

Note: results have been sorted in terms of less incorrectly clustered bigger likelihood and the number of clusters.

In ***Table 11,*** how ever the algorithms like EM(real value), Make density based cluster (real value and difference) gave the promising result, since the EM has been chosen I only show their difference between their clusters. For the rest of algorithms the same differences have been seen, so it is not necessary to show the every single one. The other algorithms like Hierarchical (Complete linkage ,Euclidean) by HBC or Farthest First or Cobweb, K-means… ,which have no promising cluster it is obvious ,do no need to be demonstrated.

The knowledge representations of the promising cluster (cluster 0) of chosen algorithm (EM with differences) from table 11 have shown in three respects:

1. Graph

Graph below shows the cluster 0 changes of (EM with differences) according to table 11. As we see in day stay 16, there is a big sudden change to decreasing trend and in day stay 17; it jumps to increasing trend (two major fluctuations).

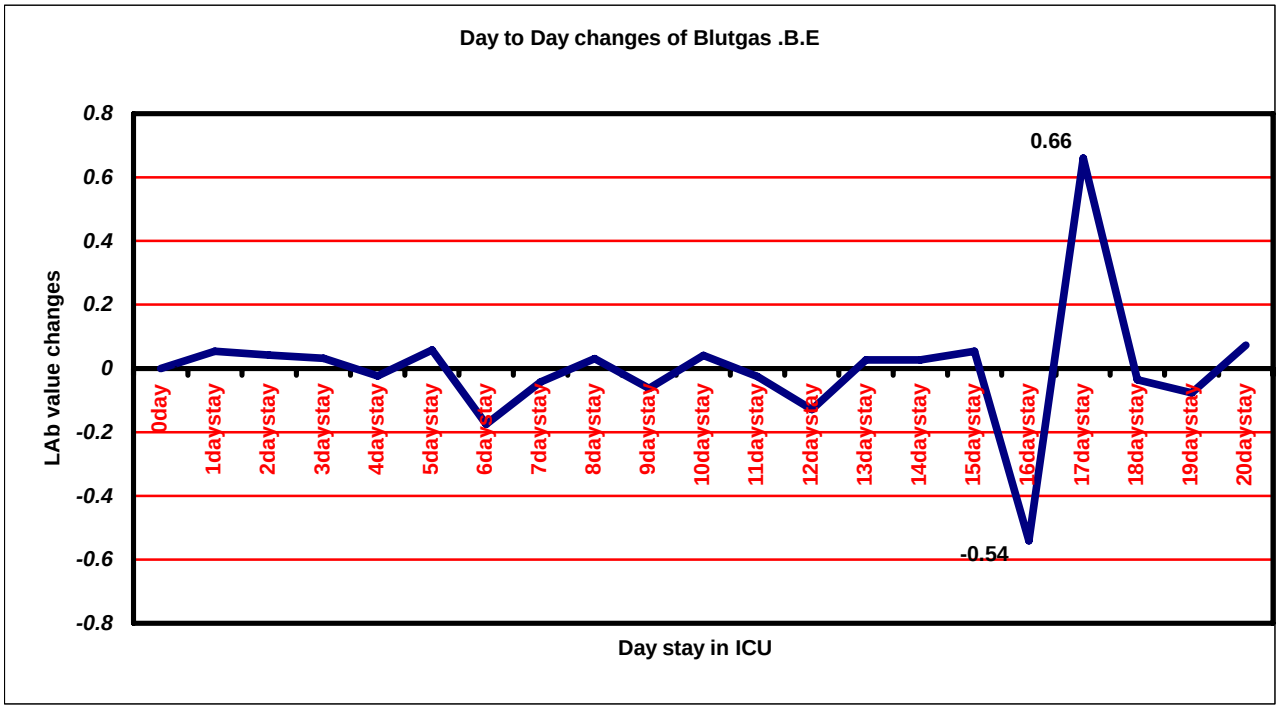

Figure 6: Presentation of changes in promising cluster achieved by EM using Δ.

2. Table of difference between promising cluster and other clusters

Table 12 shows the difference between cluster 0 (promising cluster, highlighted below) and cluster 1 of (EM with differences) chosen according to table 11.Cluster 0 is the promising cluster.

Cluster numbers	Pneumonia	Non-pneumonia	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD			
0	28	6	Daystay14-Daystay13	Daystay15 – Daystay14	Daystay16 –Daystay 15	Day stay 17 – Daystay 16
			0.02±0.07	0.05±0.01	-0.54±0.06 From day 14 to day 15 minor change but in day 16 one decreasing jump and in day 17 one increasing jump	0.66±0.08 The difference of day stayK+1 and Day stay K stored in daystayK+1
1	56	78	-0.04±0.2	-0.06±1.3 (minor change)	-0.05±0.1 (minor change)	0.04±0.9 (minor change)

Table 12: the difference between two clusters achieved using EM using Δ (best result) of Blutgas B.E.

3. Pneumonia day's match [11]

According to the study, day before or after in that pneumonia diagnosed are different from case to case. For example if one patient lab value is studied in day stay K, this day can move from -5 days of pneumonia diagnosed to 5 days after.

Thus, it is not possible to say that a pneumonia day belongs to which "day stay" In ICU. But there is an interval that we have obtained, 2 days before to 2 days after pneumonia diagnosed in which our critical day (days from 14 day stay to 17 day stay) can move .

In Figure 12: It has shown one pneumonia patient as a sample from promising cluster that present the day stay 16 in ICU is match with 1 day after pneumonia diagnosed .

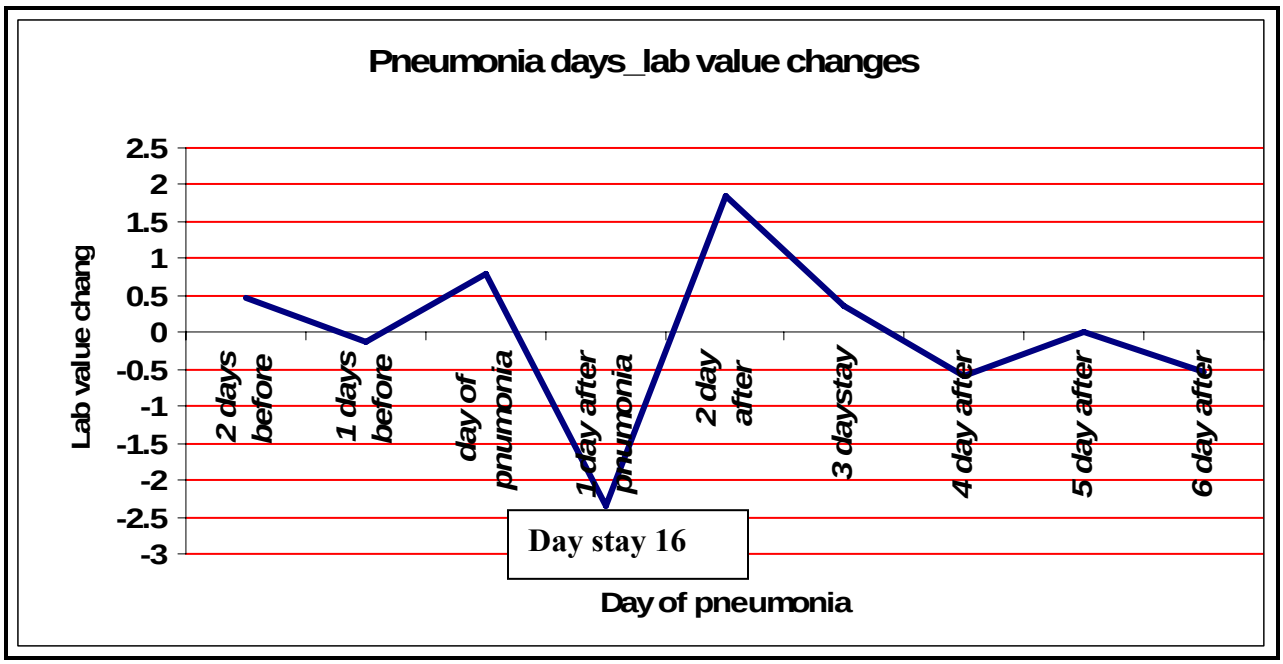

Figure 7: a pneumonia patients from Promising cluster presenting Day stay in ICU and Pneumonia day (Blutgas.B.E).

4.1.6. Conclusion of the results in Blutgas .BE.

By looking back to the previous chapters we got these results:

What ever in Section "Some hypothesis and assumption from 3.1.1." has been assumed has been experienced till here. First assumption is true promising cluster has the less percentage of whole sample, as we see in table 11 in cluster0 from EM (difference).Second assumption is true which has been shown in table 11 that using differences (Δ) gave the better results in terms of less incorrectly cluster and bigger log likelihood. Third assumption is also true, the changes and differences in two clusters are not too big, as we look at tables 7, 10, 12.

Now, the answers to the questions in **introduction**:

1-After clustering, any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why? Yes it has been achieved by using EM and using differences among attributes (day stay in ICU).

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

Yes, these important attributes are called critical days: these **critical days (day stay 14, 16, 17)** have been seen among the important attributes which made separation in clusters in table 7 and table 10 also .In another word these attributes have shown the significant changes in separate clustering of pneumonia and non-pneumonia . These **critical days** are moving in the interval of [**2 days before pneumonia diagnosed, 2 days after pneumonia diagnosed]** in pneumonia division of promising cluster (culster0 from EM (differences)).

4.2 Profile B. Leukozyten

It is translated to **Leukocyte**.

White **blood cells** or **leukocytes** are cells which form a component of the blood .The blood cells that engulf and digest bacteria and fungi; an important part of the body's defense system. Its measuring unit is Gpt/l(**Gigapartikel pro Liter**).

4.2.1 Preliminary analyses of Profile B. Leukozyten of Frequency lab value Pneumonia

Goal

Here the goal is as the same as what was mentioned in part **"**4.1.1 Preliminary analyses of Blutgas.BE**"**goal. (To see complete tables and details in Appendix B.Tables 7, 8)
We will see some difference in the frequency order between pneumonia and non pneumonia. It shows that the most frequent lab values are rather similar.

Observations

It shows that the most frequent lab values are rather similar specially the first four, some difference in the frequency order (different positions in these two tables) between pneumonia and non pneumonia but according to the below Figures 8, 9 we can conclude that both pneumonia and non-pneumonia have the same frequency values .In both pneumonia and non- pneumonia, the most frequent lab values are between 7 to 15 with Cumulative Percent more than 66%(Cumulative percentage of lab value distribution), therefore in spite of having some changes in the most frequent lab value order pneumonia and non-pneumonias' lab values expect to have a close relation.
are showing the percentage of distribution of lab values respectively in pneumonia and non-pneumonia. (Cumulative percentage of lab value distribution)

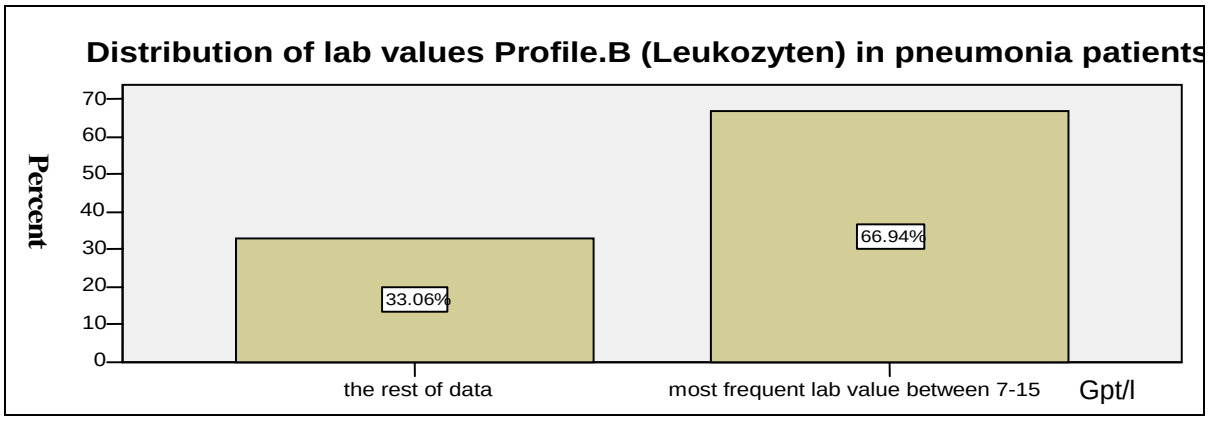

Figure 8: Distribution of lab value in pneumonia of ProfileB.Leukozyten

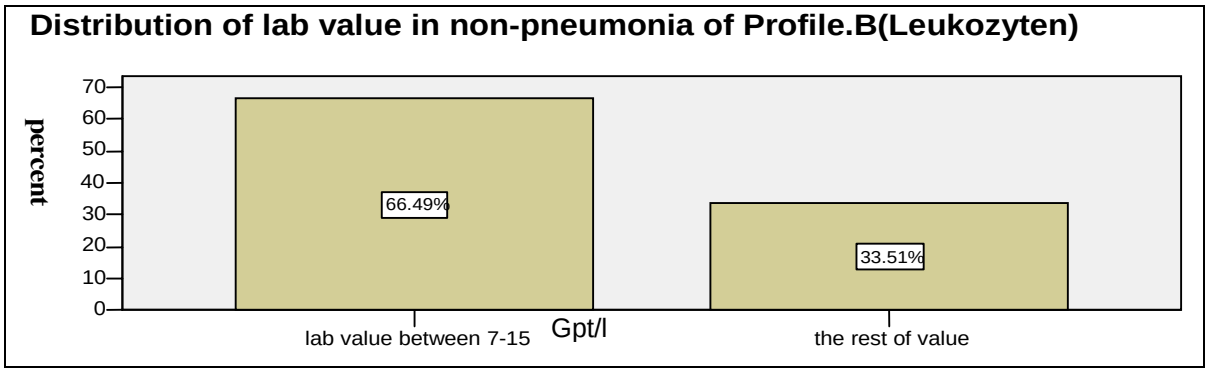

Figure 9: Distribution of lab value in non-pneumonia of ProfileB.Leukozyten

Some hypothesis and assumption from 4.2.1.

1- We can expect after clustering (joining sample) more than 60% of data will form one cluster and the rest form another cluster. In other word, promising cluster has the less percentage of whole sample.

2- Since values are too close ,for finding promising and best cluster it may be better to apply differences (Δ) also (difference between two day stay e.g.daystay (k)-daystay (k+1)) to see If it helps to get better results.

3- we can expect that ,in clustering ,probably only minor differences will may seen to make a difference between promising cluster and other clusters.

4.2.2 Preliminary analyses of Profile B. Leukozyten of out range of medical:

The medical range that was inserted for this lab values are as **(3.8 9.8)** (Gpt/l)**.**

Technique

The same technique that has been used in 4.1.2 is used.

Results

By comparing tables 13 and 14, the number of patients and their percentage of number of patients going out of range of Maximum of medical range is more probable, because the % number of patients that their labs values are out range of (9.8) that is in the range 10 to 100% are 91% in pneumonia and 84% in non-pneumonia .In addition in outrange of Minimum the percentage number of patients less than 10% in pneumonia is 96% and in non-pneumonia is 97% Therefore, we can conclude that both pneumonia and non- pneumonia cases have the tendency of going out range of maximum medical value (9.8),. So it was decided to replace the multi value in one day with maximum, which after was found that it would not give any good result in clustering comparing with when we used average. These results are showing again how much **lab values in pneumonia and non-pneumonia are close** and only minor differences that make a difference between promising cluster and other clusters.(for complete tables, **see appendix B table 7.1,7.2**).

Percentage of outrange (3.8)min medical of lab value Profile B. Leukozyten	Number of patients in pneumonia (among 93 patients)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 1001)	% number of patients in non-pneumonia
0-10%	89	96%	973	97%
10-100%	4	4%	28	3%

Table 13.Statistics value of Profile B. Leukozyten *out range (3.8) in range (***3.8 9.8***) (Pneumonia and non-pneumonia patients)*

Percentage of outrange (9.8)max medical of lab value Profile B. Leukozyten	Number of patients in pneumonia (among 93 patients)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 1001)	% number of patients in non-pneumonia
0-10%	8	8.6%	158	16%
10-100%	85	91.4%	843	84%

*Table 14.Statistics value of Profile B. Leukocyte range (9.8) in range (***3.8 9.8***). (Pneumonia and non-pneumonia patients)*

4.2.3 Preparation stage of Profile B. Leukocyte by separate clustering Pneumonia: [18]

a) Pneumonia (average, real value)

This section has been skipped because cluster results does not show any difference between classes achieved (only minor change).

b) Pneumonia (Δ, difference of data average)

The approach was started by clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) .Thus, what can be done is to see how the changes are, so instead of studying lab value number itself, we calculated the differences (Δ) day to day and then try to cluster them again

The 74 attributes are (age, sex, 0daystay, 1daystay… 71daystay) used EM algorithm.

The result of clustering

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	63	-47.11
1	10	
2	18	

Table 15: Clustering results of non- pneumonia in Profile B. Leukocyte using differences (Δ) day to day.

Difference of clusters

Table 16 shows the difference among the 3 clusters that have been achieved above.

This clustering result is done by EM.The interesting differences happened in difference of Day stay 13 in cluster 1 is increasing while the same day in other clusters are decreasing. It is observed that in cluster 2 Day stay 12 is increasing while in other clusters it is decreasing.

Cluster numbers	Number of member	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD		
		Daystay11-Daystay10	Daystay12-Daystay 11	Daystay13-Daystay12
0	63	0.8±0.2	-0.17±0.06	-0.29±0.14
1	10	0.2±0.1	-0.2 ± 0.09	0.6±0.4
2	18	0.2± 0.009	1.9±0.76	-0.77±0.8)

*Table16: Summary of clustering all pneumonia of lab value Profile B. Leukocyte and its clusters Differences using difference of mean value of data day to day (*Δ*).*

4.2.4 Preparation stage of Profile B. Leukozyten by separate clustering Non-Pneumonia

a) Non-Pneumonia (average, real value)

This section has been skipped because cluster results do not show significant result comparing with results achieved by using the differences of lab values (day to day) (only minor change.

b) Non-Pneumonia (Δ, difference of days)

The approach was started by clustering all pneumonia patients, the rest are the same as the section 4.2.3.b.

The result of clustering

clusters	Number of members in cluster(out of 126 cases)	Log likelihood
0	48	-17.718
1	107	
2	182	
3	647	

Table 17: Clustering results of non- pneumonia in Profile B. Leukocyte using differences (Δ) day to day.

Difference of clusters

Table 16 shows the difference among the 3 clusters that have been achieved above.

This clustering result is done by EM.The interesting differences happened in difference of Day stay 13 in cluster 1 is increasing while the same day in other clusters are decreasing.

It is observed that in cluster 2 Day stay 12 is increasing while in other clusters it is decreasing.

Note :(The difference (daystayK+1 –Day stay K) is stored in daystayK+1 **)**

Cluster numbers	**Number of member**	**Comparison of Cluster s by Δ** **mean (daystayK+1 –Day stay K) ± STD**		
		Daystay11-Daystay 10	**Daystay12-Daystay 11**	**Day stay 13-Daystay12**
0	**48**	**-0.2 ±0.01**	**-0.3±0.05**	**-0.02±0.01**
1	**107**	**0.2±0.01**	**0.9±0.02** (As we see day12 are different in this cluster with others ,there is a increasing trend and others are decreasing and also in Day stay 13 there is a sudden change and is bigger than other cluster in this day)	**-0.88±0.01**
2	**182**	**-0.5±0.011**	**-0.5±0.4**	**-0.04± 0.5**
3	**18**	**0.02±0.6**	**-0.4±0..03**	**-0.01±0.02**

Table18: Summary of clustering all pneumonia of lab value Profile B. Leukozyten *and its clusters differences using difference of mean value of data day to day* Δ

4.2.5 Clustering Joining Pneumonia and Non-pneumonia Profile B. Leukozyten:

Technique

The same technique that mentioned in 4.1.5 for Blutgas .BE. has been used here also.

Results

In Table 19 results have been sorted in terms of less incorrectly clustered and then bigger likelihood, and then number of clusters .The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

What has to be mentioned about Table ***19*** that it shows the difference between applying different algorithms. **The reason that the result from EM using differences has been chosen** are as follows:

1-The distribution of data in cluster 1(promising one) looks reasonable 17% of pneumonia and 5 % non- pneumonia (15, 5). (**A cluster with majority is with pneumonia**).

2- **The percentage of incorrectly clusters instances are less than other algorithms**.

3-Comparing with other algorithms like EM(real value),K- Make density based cluster(real value and difference) ,which are having a promising cluster, **it has the better log likelihood function** (since likelihood measures how likely clustering is, so the greater the log likelihood is, the better the clustering result is).

Algorithm	clusters	Pneumonia (91 patients)	Non-pneumonia (91patients)	Percentage Of total sample	Data	**Log likelihood**	**Incorrectly clustered instances**	Number of clusters achieved
EM	0	4	7	6%	Difference	-37.2	**42%**	3
	1	15	5	10%				
	2	72	79	84%				
Make density based cluster	0	8	4	7%	Difference	-52.4	**46%**	2
	1	83	87	93%				
K-means	0	61	55	64%	Difference	----------	**46%**	2
	1	30	36	36%				
K-means	0	54	61	63%	Real	-----	**47%**	2
	1	37	30	37%				
Make density based cluster	0	75	73	81%	Real	-137.09	**48%**	2
	1	16	18	19%				
EM	0	1	0	1%	Real	-72.7	**49%**	4
	1	12	8	11%				
	2	9	3	7%				
	3	69	80	82%				
Farthest first	0	90	91	99%	Real	----------	**49%**	2
	1	1	0	1%				
Farthest first	0	90	91	99%	Difference	----------	**50%**	2
	1	1	0	1%				
Hierarchical (Complete linkage ,Euclidean)	0	4	13	9%	Difference	----------	----------	2
	1	87	78	91%				
Hierarchical (Complete linkage ,Euclidean)	0	91	91	--	Real	----------	----------	1
cobweb	none				Real		**96%**	271
cobweb	none				Difference		**98%**	255

Table 19**:** *difference of applying different algorithm for clustering using Weka and HCE. Profile B.Leukozyten.*

The knowledge representations of the promising cluster (cluster 1) of chosen algorithm (EM with differences) from table 19 have shown in three respects:

1. Graph

Graph below shows the cluster 1 changes (from EM with differences) according to table 19.

As we see in day stay 3, there is a decreasing trend and suddenly increase in day stay 4 which are significant compare with other changes. In addition, there is decreasing and sudden increasing jump in day stay 12 and suddenly decrease jump in day stay 13. **Two days are the most critical points, day stay 3 and daystay12 (**the one which has shown significant changes in table 16, 18**).**

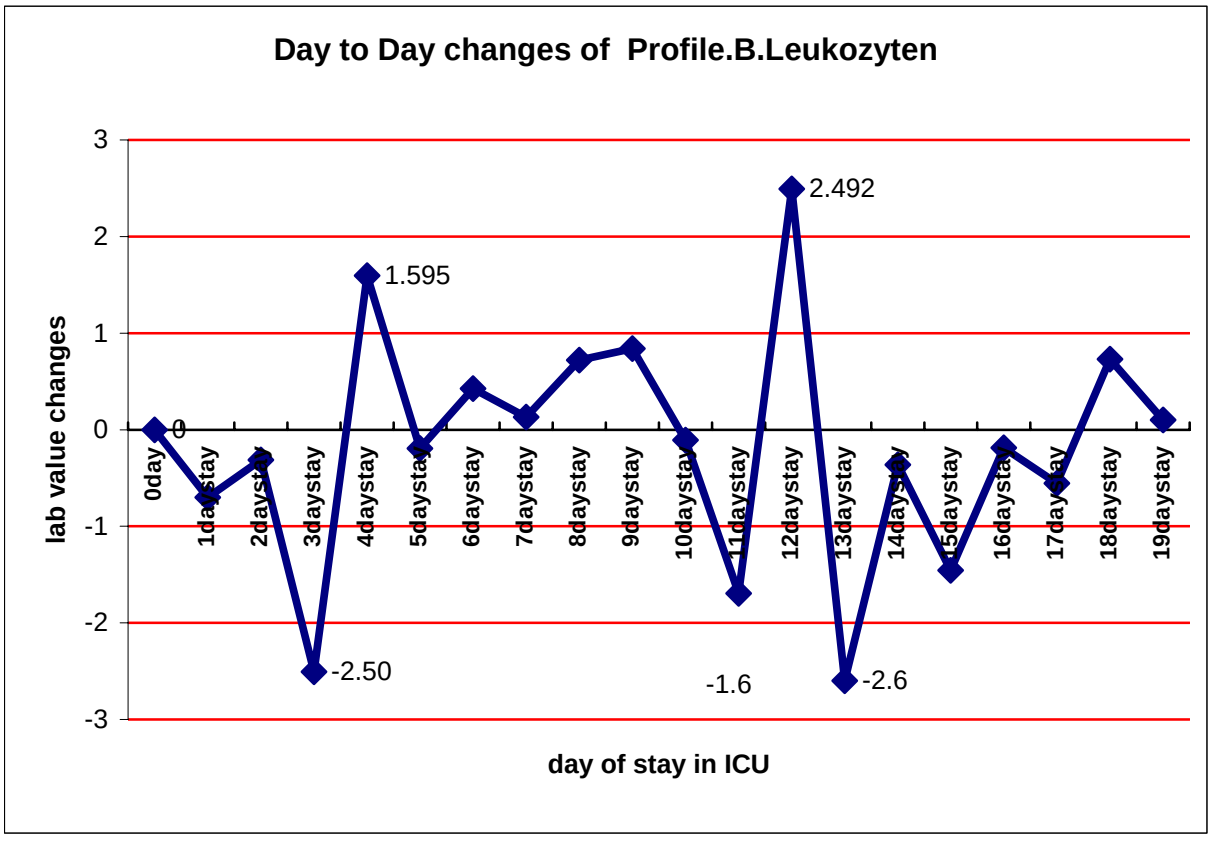

Figure 10: Lab value changes trend of Profile B. Leukozyten.

2. Table of difference between promising cluster(1) and other clusters

Table 20 shows the difference between cluster 1(promising cluster, highlighted below) and cluster 1 (used EM algorithm with differences).

Cluster numbers	Pneumonia	Non-pnumonia	Comparison of Cluster s by Δ mean (day stay(K+1) –Day stay K) ± STD [19,21] The difference of daystayK+1 –Day stay K stored in day stay k+1				
			Day stay 3-Day stay2	Day stay 4-Day stay 3	Day stay 11-Day stay 10	Day stay 12-Day stay 11	Daystay13-Daystay12
0	7	4	0.4±0.2	0.4±0.01	0.5±0.3	0.6±0.07	-0.5±0.06
1	15	6	-2.5±0.3	1.5±0.6	-1.6± 0.05	2.5±0.4	-2.6±0.3
2	72	79	-0.4±0.03	-1.02±0.23	0.2±0.07	0.2±0.05	-0.7±0.05

Table 20: the difference between two clusters achieved using EM using Δ (best result) of Profile B. Leukozyten.

3. Pneumonia days match [11]

It is not possible to say that a pneumonia day belongs to which "day stay" In ICU (more explanations .see "3. Pneumonia days match in 4.1.5".

There are two different intervals have been found, by looking at the only pneumonia patients of cluster 1 .First Interval is for the first critical points Day stay 3 and Day stay 4 ,which moves in interval [-5 ,3] =[5 days before pneumonia,3 day after pneumonia].

The Second Interval belongs to the second critical points Day stay 11 to

Day stay 13, which moves in interval [-5, 5] = [5 days before pneumonia, 5 day after pneumonia].

Thus, in terms of pneumonia view we can say in overall two fluctuations are in the interval [-5 5].

In Figure 11 has shown one pneumonia patient as a sample from promising cluster(1) that present the day stay 3 in ICU matches with 4 day after pneumonia diagnosed and Day stay 12 matches with day stay 5.

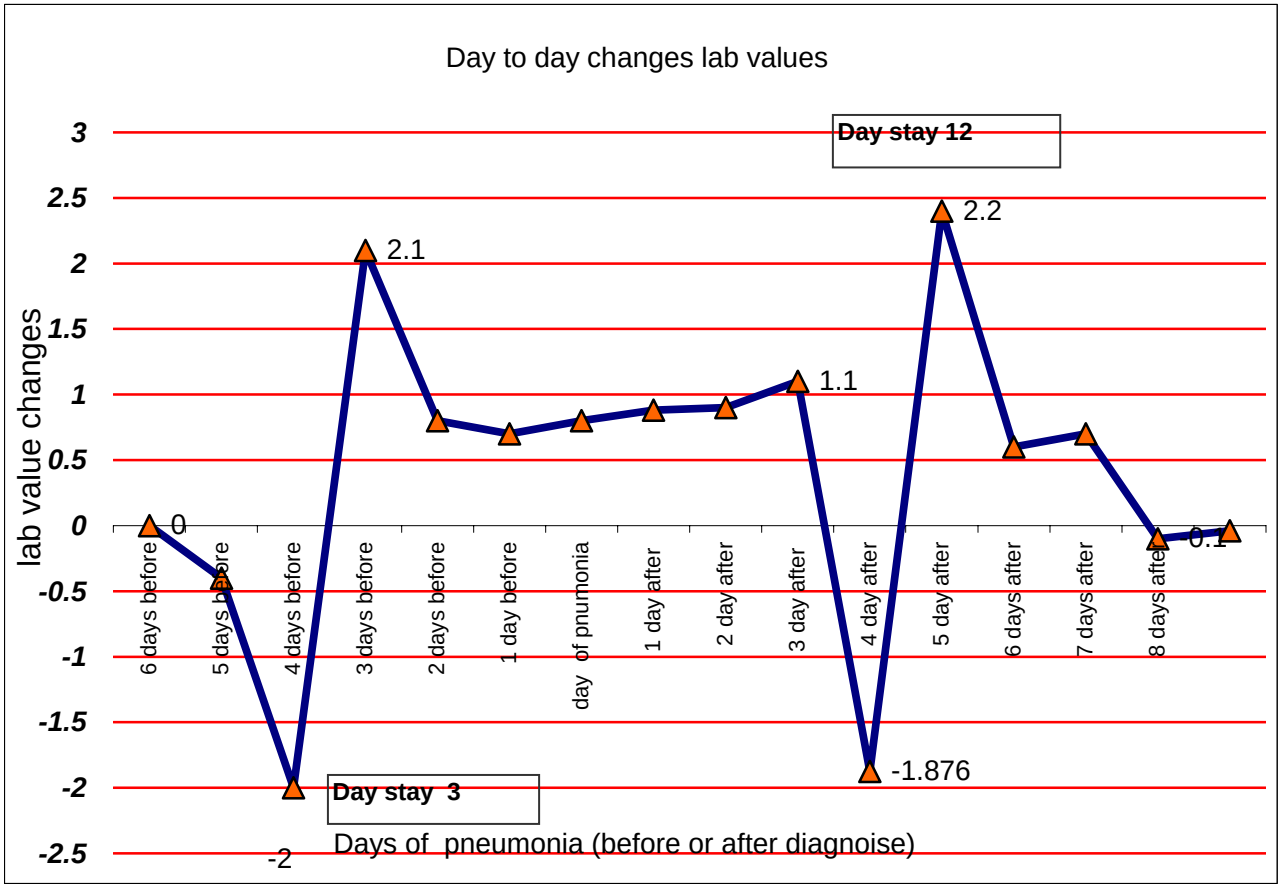

Figure 11: a pneumonia patients from Promising cluster presenting Day stay in ICU and Pneumonia day. (Profile B.Leukozyten)

4.2.6. Conclusion of the results in Profile B.Leukozyten.

By looking back to the previous chapters we got these results:

What ever in Section "Some hypothesis and assumption from 4.2.1." has been assumed **has been** experienced. First assumption is true promising cluster has the less percentage of whole sample, as we see in table 19 in cluster1 from EM (difference).Second assumption is true which has been shown in table 19 that using differences (Δ) gave the better results in terms of less incorrectly cluster and bigger log likelihood. Third assumption is also true, the changes and differences in two clusters are not too big, but this time, comparing with blutgas.B.E are more reasonable.

Now, the answers to the questions in **introduction**:

1-After clustering, any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why? Yes it has been achieved by using EM and using differences among attributes (day stay in ICU).

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

Yes, these important attributes are called critical days: these critical days are day stay 3, day stay 4,day stay 11,day stay 12,day stay 13 which made separation in clusters in table 20 . These **critical days** are moving in the interval of [**5 days before pneumonia diagnosed, 5 days after pneumonia diagnosed]** in pneumonia division of promising cluster (culster1 from EM (differences)).

4.3 Blutgas arteriell .HB

It is translated to **Arterial Blood Gas** which is White blood cells and Hb is Hemoglobin of blood.

4.3.1 Preliminary analyses of Blutgas.HB of Frequency lab value

Goal

Here the goal is as the same as what was mentioned in part **"**4.1.1 goal. (To see complete tables and details in Appendix B.Tables9, 10)

It shows that the most frequent lab values are similar in pneumonia and non-pneumonia.

Observations

It shows that the most frequent lab values are rather similar specially the first four, in pneumonia and non pneumonia .According to the below Figures 12, 13 we can conclude that both pneumonia and non-pneumonia have the same frequency values .In both pneumonia and non- pneumonia, the most frequent lab values are between 4 to 8 with Cumulative Percent (Cumulative percentage of lab value distribution) more than 96%. So Lab values in pneumonia and non-pneumonia have a tight relation.

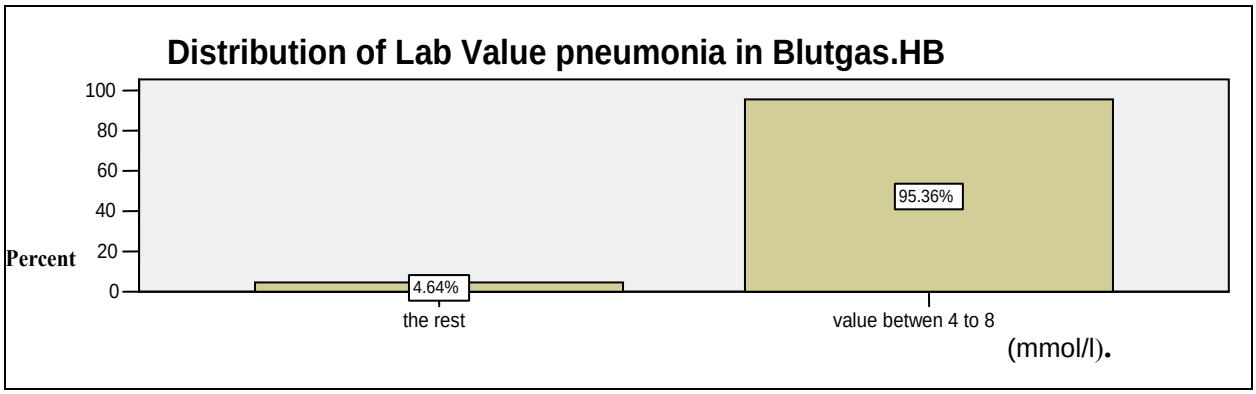

Figure 12: Distribution of lab value in pneumonia of Blutgas.HB.

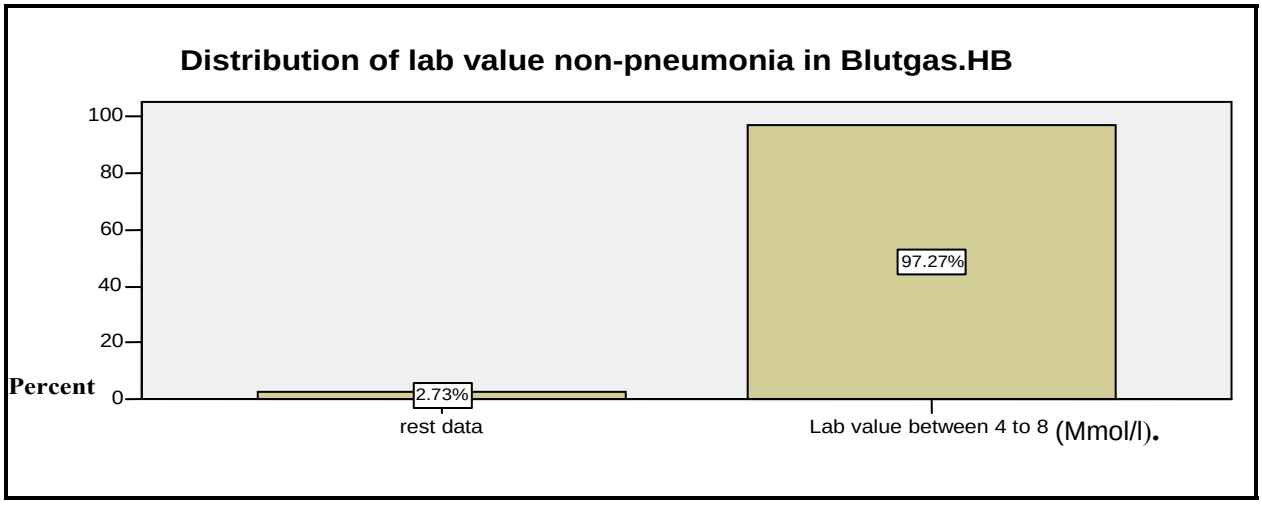

Figure 13: Distribution of lab value in non-pneumonia of Blutgas.HB

.

Some hypothesis and assumption from 4.3.1.

1- We can expect after clustering (joining sample) promising cluster has the less percentage of whole sample.

2- Since values are too close ,for finding promising and best cluster it may be better to apply differences (Δ) also (difference between two day stay e.g. day stay (k)-day stay (k+1)) to see If it helps to get better results.

3- We can expect that ,in clustering ,probably only minor differences will may seen to make a difference between promising cluster and other clusters.

4.3.2 Preliminary analyses of Blutgas.HB of out range of medical:

The medical range that was inserted for this lab values are as **(8.6, 12).** (Mmol/l)

Technique

The same technique that has been used in 4.1.2 is used and following results have been achieved:

Results

By comparing tables 21 and 22, the percentage of number of patients that their lab values going out of range of Minimum of medical range has a big percentage. % number of patients that their labs values are out range of (8.6) that is in the range 10 to 100% are 100% in pneumonia and 98% in non-pneumonia .In addition in outrange of Maximum the percentage number of patients in both pneumonia and non-pneumonia that their lab values goes out of (12) is zero in table 22.

Therefore, we can conclude that both pneumonia and no pneumonia cases have the tendency of going out range of minimum (6.8) in medical range (6.8, 12). So this is used to replace the multi value in one day with minimum, which did not give good result in clustering comparing with when we used average. These results are showing again how much **lab values in pneumonia and non-pneumonia are close** and only minor differences that make a difference between promising cluster and other clusters. (For complete table, **see appendix B table 11**).

Percentage of outrange (8.6)min medical of lab value *Blutgas.HB*	Number of patients in pneumonia (among 92 patients)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 813)	% number of patients in non-pneumonia
0-10%	0	0%	16	2%
10-100%	92	100%	797	98%

Table 21.Statistics value Blutgas.HB out range (3.8) in range **(8.6, 12)***(Pneumonia and non-pneumonia patients)*

Percentage of outrange (12)max medical of lab value *Blutgas.HB*	Number of patients in pneumonia (among 92 patients)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 813)	% number of patients in non-pneumonia
0-10%	0	0%	0%	0%
10-100%	0	0%	0%	0%

Table 22...Statistics value of Blutgas.HB range (12) in range **(8.6, 12).** *(Pneumonia and non-pneumonia patients)*

4.3.3 Preparation stage of Blutgas.HB by separate clustering Pneumonia: [18]

a) Pneumonia (average, real value)

It is clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) to see if there is any attributes that show a big difference among clusters?

The Attributes are used for clustering are: 74 including (age, sex, 0daystay, 1 day stay… 71daystay).

The results achieved by using EM algorithm are as follows:

clusters	Number of members in cluster(out of 88 cases)	Log likelihood
0	64	13.47
1	24	

Table 23: Clustering results of pneumonia in Blutgas.HB .using real values.

Cluster differences

Table below shows the difference between clusters achieved in this clustering by EM.

The interesting differences(attributes that made the differences) happened from Day stay 0 to Day stay 3 in cluster 1 .In these days value is in range 7(7<=X<8)in cluster 1, while in Cluster 0 are in range 5(5<=X<6).

Cluster numbers	**Number of member**	**Differences between Cluster 0 and cluster 1 by comparisons of average (μ± STD)** [19][21]			
		Day stay 0	**Day stay 1**	**Day stay 2**	**Day stay 3**
0	**64**	**6±0.08**	**5.9±0.06**	**5.8±0.5**	**5.9±0.4**
1	**24**	**7.26±0.2**	**7.61±0.09**	**7.2±0.09**	**7.2±0.08**

Table 24: Summary of clustering all pneumonia of lab value Blutgas.HB and its clusters differences using mean value data.

b) Pneumonia (Δ, difference of data average)

This section has been skipped because cluster section "a. Pneumonia (average, real value)" can easily see the difference among cluster.

4.3.4 Preparation stage of *Blutgas.HB* by separate clustering Non-pneumonia:

a) Non-Pneumonia (average, real value)

The same approach that mentioned in "a. Pneumonia (average, real value)", Results are:

clusters	Number of members in cluster(out of 88 cases)	Log likelihood
0	40	31.65
1	773	

Table 25: Clustering results of all non-pneumonia in Blutgas.HB .using real values.

Cluster differences

The same attributes (day stay 0 to day stay 5) that have been achieved in Pneumonia (average, real value)" have shown the differences among two clusters compare with other attributes.

Cluster numbers	**Number of member**	**Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)**			
0	41	**Day stay 0**	**Day stay 1**	**Day stay 2**	**Day stay 3**
		6.9±0.09	**7.8±0.08**	**8.3±0.07**	**6.88±0.06**
1	772	**5.3±0.3**	**7.0±0.03**	**5.9±0.02**	**5.01±0.02**

Table 26: Summary of clustering all non-pneumonia of lab value Blutgas.HB and its clusters differences using mean value data.

b) Non-Pneumonia (Δ, difference of data average)

Here skipped to show because the better results achieved by using real value.

4.3.5 Clustering Joining Pneumonia and Non-Pneumonia *Blutgas.HB:*

Technique

The same technique that mentioned in 4.1.5 for Blutgas .BE. has been used here.

The sample includes (50% pneumonia and 50% non-pneumonia) and different clustering algorithm has been used.

Results

In Table 19 results have been sorted in terms of less incorrectly clustered and then bigger likelihood. The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

What has to be mentioned about Table ***27 is*** that shows the difference between applying different algorithms. **The reason that the result from EM using real value has been chosen** are as follows:

1-The distribution of data in cluster 0(promising one) looks reasonable 55% of pneumonia(49 out of 88 patients) and 20% non pneumonia (18 out of 88 patients).(**A cluster with majority is with pneumonia**).

2- **The percentage of incorrectly clusters instances are less than other algorithms**.

3-Comparing with other algorithms like Make density based cluster(real value) ,which are having a promising cluster, **it has the better log likelihood function** (since likelihood measures how likely clustering is, so the greater the log likelihood is, the better the clustering result is).

Algorithm	Clusters	Pneumonia 88 patients	Non-pnu 88 patients	percentage	Data	Log likelihood	**Incorrectly clustered instances**	Number of clusters achieved
EM	0	49	18	40%	real	139.25	**35%**	2
	1	39	70	60%				
Make density based cluster	0	24	25	27%	real	68.9	**45%**	3
	1	3	1	4%				
	2	61	62	69%				
K-means	0	12	30	24%	real	-----	**48%**	2
	1	76	58	76%				
Farthest first	0	88	87	99%	real		**49%**	2
	1	0	1	1%				
Make density based cluster	0	88	88	100 %	difference	167	**49%**	1
EM	0	88	88	100 %	difference	166	**49%**	1
K-means	0	88	88	100 %	difference	---	**49%**	1
Farthest first	0	88	88	100 %	difference		**50%**	1
cobweb	none				real	--	**89%**	118
cobweb	none				difference	---	**89%**	146
Hierarchical (Complete linkage ,Euclidean)	0	88	88	100 %	real	--	**--**	1
Hierarchical(Complete linkage,Euclidean)	0	88	88	100 %	difference	----	**---**	1

Table 27: difference of applying different algorithm for clustering using Weka and HBC Blutgas.HB

The knowledge representations of the promising cluster (cluster 0) of chosen algorithm (EM with real value) from table 27 have shown in three respects:

1. Graph

Graph below shows the cluster 1 changes:

The first 4 days stay in ICU the lab value are in the range of 7 and for next days is started to decreasing to range 6 and then rang 5,so From day stay 0 to day stay 4 are the critical days which made the difference among cluster 1 and cluster 0.

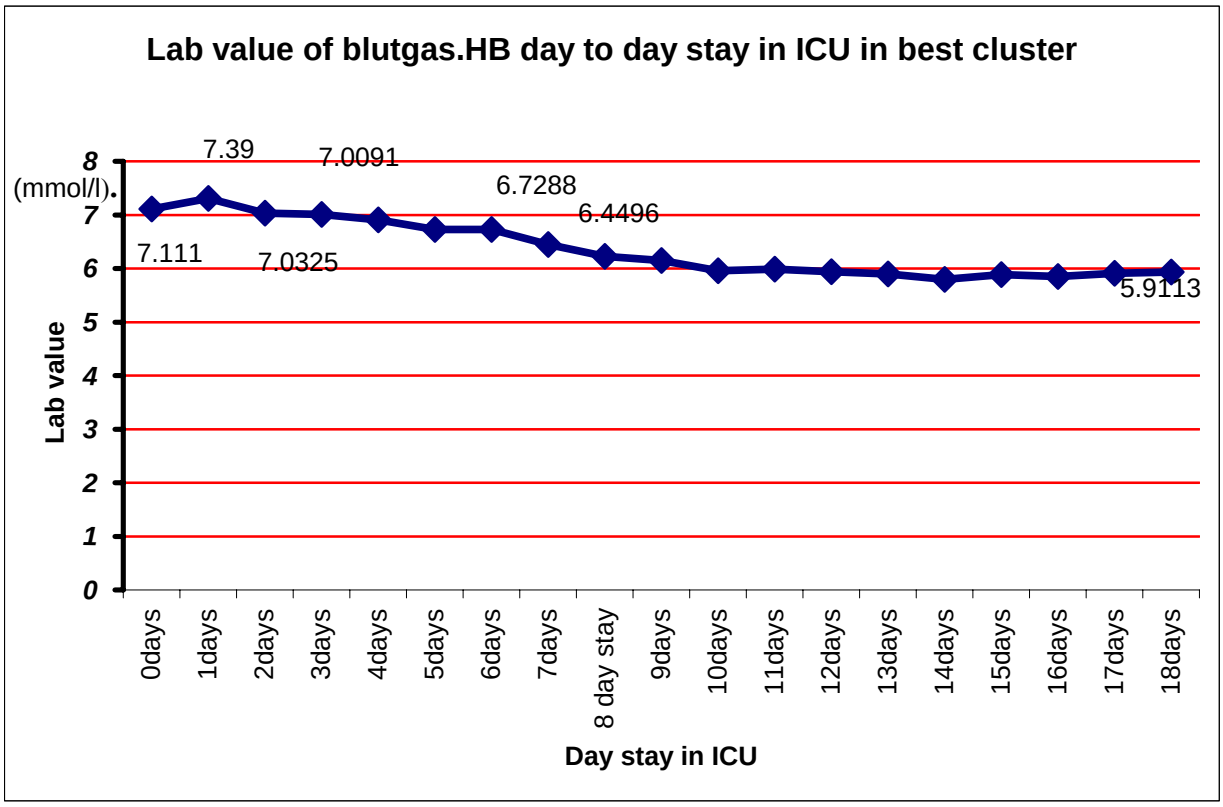

Figure 14: Lab value day to day changes trend in ICU of Blutgas.HB

2. Table of difference between promising cluster(1) and other clusters

Table 28 shows the difference between cluster 1(promising cluster, highlighted below) and cluster 1 .

Cluster	Number of pneumonia	Number of non-pneumonia	**Differences between Cluster 0 and cluster 1 by comparisons of average** (value average ± STD)				
			Day stay 0	Day stay 1	Day stay 2	Day stay 3	Day stay 4
0	49	18	7.1±0.09	7.3±0.09	7.03±0.08	7.009±0.08	7±0.7
1	39	70	5.9±0.05	5.8±0.07	5.7±0.07	5.7±0.07	5.8±0.08

Table 28: the difference between two clusters achieved using EM (real value)of Blutgas.HB.

3. Pneumonia days match [11]

The critical days from 0 day stay to 4th day stay can be vary between 4 days before pneumonia to 4 days after pneumonia happened [-4,4] ,So we can say the first 4 days stay in ICU, are changing in the interval of -4 to 4days of pneumonia has been diagnosed. In Figure 15, one pneumonia patient as a sample has been shown to present the pneumonia days matches with which day stay in ICU.

Here, pneumonia day (0) is match with day stay 1 and 1 day after pneumonia diagnosed is match with day stay 2, and so on.

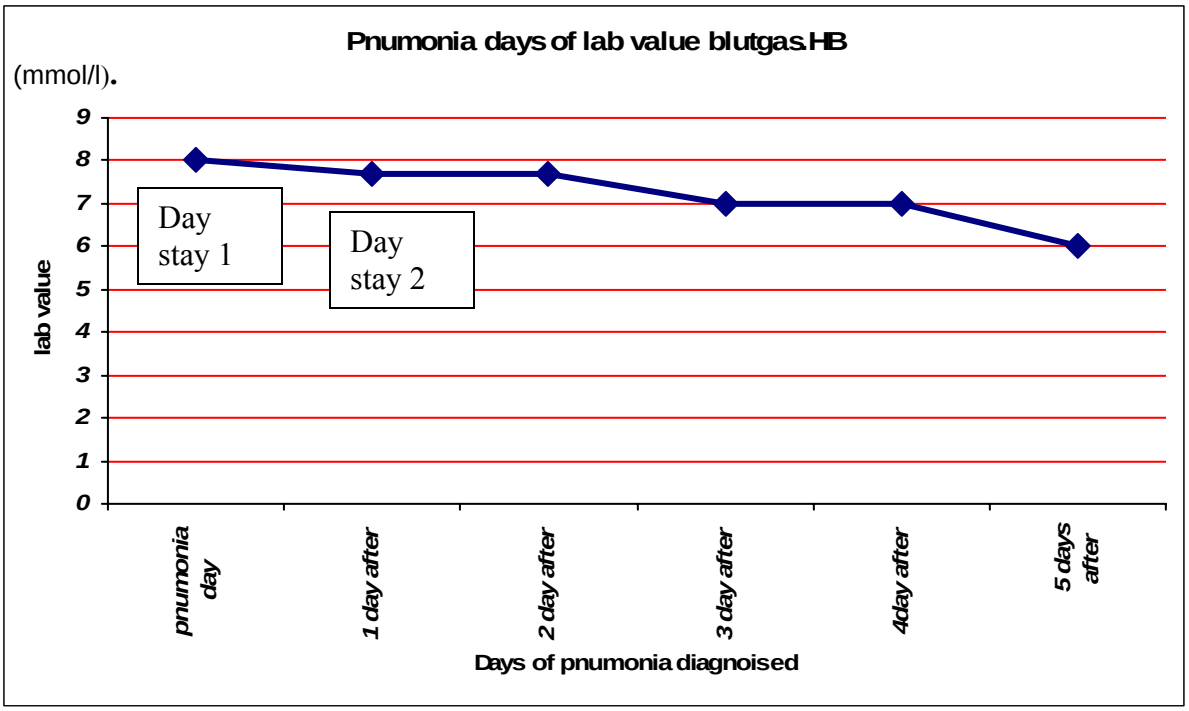

Figure 15: a pneumonia patients from Promising cluster presenting Day stay in ICU and Pneumonia day. (Blutgas.HB).

4.3.6. Conclusion of the results in Blutgas.HB

By looking back to the previous chapters we got these results:

What ever in Section "Some hypothesis and assumption from 4.3.1." has been assumed**, has been** experienced till here. First assumption is true promising cluster has the less percentage of whole sample, as we see in table 27 in cluster1 from EM (Real).Second assumption is not true which has been shown in table 27 that using differences (Δ) gave the better results in terms of less incorrectly cluster and bigger log likelihood . Third assumption is true, the changes and differences in two clusters are not too big, but this time.

Now, the answers to the questions in **introduction**:

1-After clustering, any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why? Yes it has been achieved by using EM and using Real value (value itself, not the difference among value of lab data day to day).

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

Yes, these important attributes are called critical days: these critical days are day stay 0, day stay 1,day stay 2,day stay 3,day stay 4 which made separation in clusters in table 28 . These **critical days** are moving in the interval of [**4 days before pneumonia diagnosed, 4 days after pneumonia diagnosed]** in pneumonia division of promising cluster (culster1 from EM (differences)).

Chapter 5

5.1 Blutgas arteriell .COHB

It is translated to Blood gas arterial that has COHB component which is CARBOXYHAEMOGLOBIN in blood.

5.1.1 Preliminary analyses of Blutgas. COHB of Frequency lab value:

Goal

Here the goal is as the same as what was mentioned in part **"**4.1.1 Preliminary analyses of Blutgas.BE**"**goal. (To see complete tables and details in Appendix B.Tables 12, 13)
We will see some difference in the frequency order between pneumonia and non-pneumonia. It shows that the most frequent lab values are the same.

Observations

It shows that the most frequent lab values are the same in pneumonia and non pneumonia according to the below Figures 16, 17. In pneumonia lab values are between 0 to 4 (With Cumulative Percent 100%) and the most frequent lab values in non- pneumonia is also between 0 to 4 (With 99% Cumulative Percent percentage of lab value distribution), the lab values in pneumonia and non-pneumonias' lab values have a close relation.

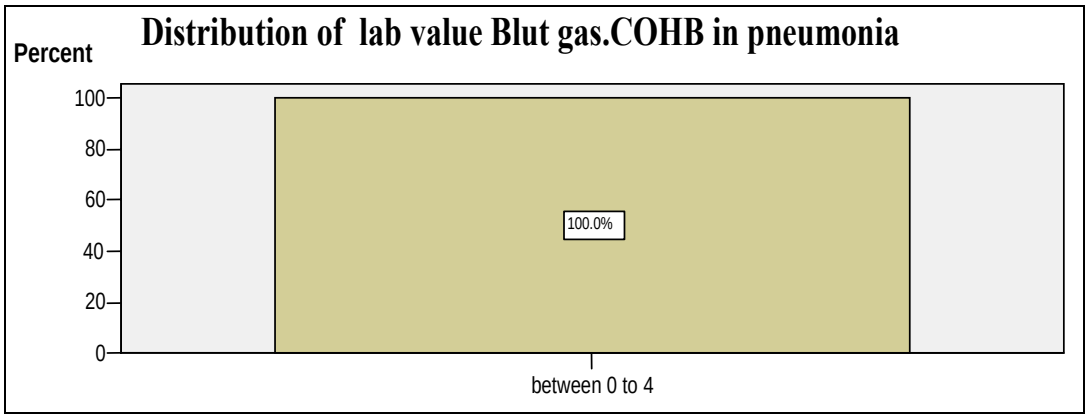

Figure16: Distribution of lab value in pneumonia of Blutgas.Cohb

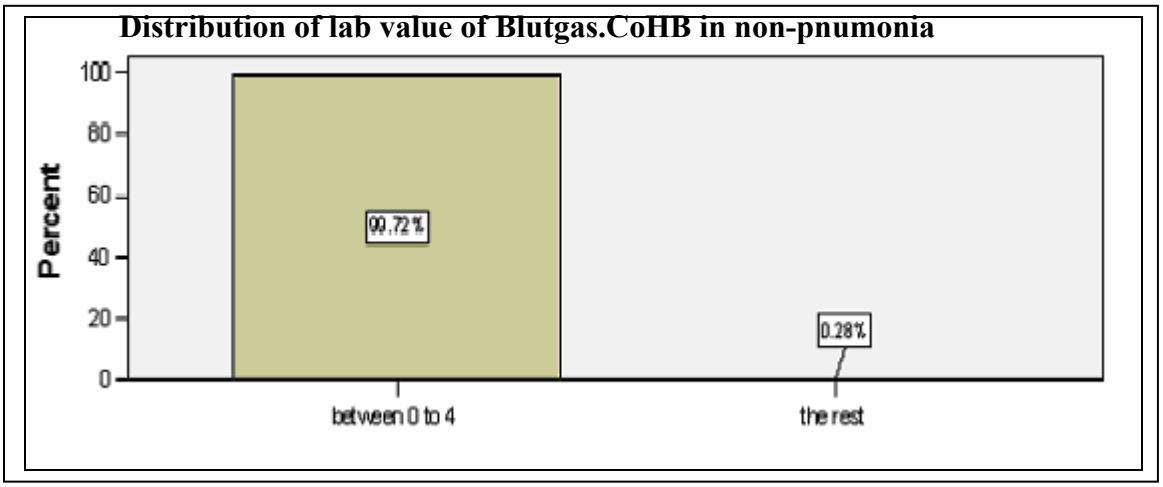

Figure 17: Distribution of lab value in non-pneumonia of Blutgas.Cohb.

Some hypothesis and assumption from 5.1.1.

1-If any promising cluster achieves it will have the less percentage of whole sample.

2- Since values are too close ,for finding promising and best cluster it may be better to apply differences (Δ) also (difference between two day stay e.g.daystay (k)-daystay (k+1)) to see If it helps to get better results.

3- we can expect that ,in clustering ,probably only minor differences will may seen to make a difference between promising cluster and other clusters.

5.1.2 Preliminary analyses of Blutgas.HB.of out range of medical:

Section 5.1.2 "Preliminary analyses of *blutgas.COHB* of out range of medical data has been skipped because, there is no medical range inserted for this lab data.

5.1.3 Preparation stage of Blutgas.COHB by separate clustering Pneumonia:

a) Pneumonia (average, real value)

It is clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) to see if there is any attributes that show a big difference among clusters?

The Attributes are used for clustering are: 74 including (age, sex, 0daystay, 1 day stay… 71daystay).

The results achieved by using EM algorithm are as follows:

Answer: NO, by studying the two different clusters, no big difference between two clusters has been found to make a decision.

Cluster numbers	Number of member	Log likelihood
0	19	166.6
1	69	

Table 29: Clustering results of pneumonia in Blutgas.CoHB .using real values.

Note: The difference among two clusters achieved above is too small .Clustering with differences shows better how these clusters are close since there is no separations achieved by using **Δ.** In part b **(the next section), See Appendix B table 14.**

b) Pneumonia (Δ, difference of data average)

The same approach as motioned in above section "a..5.1.3".

The result is:

Cluster numbers	Number of member	Log likelihood
0	**88**	180.5

Table 30: Clustering results of pneumonia in Blutgas.CoHB .using difference of data average.

As we see, no splitting of data has been seen when we use Δ.which we expected from section "a.5.1.3".

5.1.4 Preparation stage of Blutgas.COHB by separate clustering Non-Pneumonia:

a) Non-Pneumonia (average, real value)

The same approach and attributes that has been used in section a 5.1.3 used here also.

Cluster numbers	Number of member	Log likelihood
0	**784**	**195**
1	**103**	

Table 31: Clustering results of non- pneumonia in Blutgas.CoHB .using real value.

The same thing we have got in non-pneumonia patients have been achieved here, no promising or significant difference among these clusters to make us to rely on. The difference is shown below as an example:

Cluster numbers	Number of member	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)			
		Day stay 3	Day stay 4	Day stay 5	Day stay 6
0	784	1.24±0.04	1.16±0.045	1.29±0.04	1.3±0.04
1	103	1.27±0.07	1.11±0.3	1.27±0.06	1.42±0.03

Table 32: the difference between two clusters achieved by Weka for Blutgas. COHB in non-Pneumonia. using real value.

Note: As we see in table 32 the differences between two clusters in certain day are too small to rely on. For example in Day stay 3 the maximum difference are 0.05.

b) Non-Pneumonia (Δ, difference of data average)

The same approach and attributes that has been used in section a 5.1.3 used here also

Cluster numbers	Number of member	Log likelihood
0	104	169.5
1	748	
2	25	
3	10	

Table 33: Clustering results of non- pneumonia in Blutgas.CoHB .using difference of data average.

The differences are so minor and they are not big enough or significant (see appendix b table 15).Thus, it is not possible to choose any cluster which has a major or noticeable difference from others. So far no promising cluster has been found neither in Pneumonia nor in Non-pneumonia.

5.1.5 Clustering Joining Pneumonia and Non-Pneumonia *Blutgas. COHB:*

Technique

The same technique that mentioned in 4.1.5 for Blutgas .BE. has been used here. The sample includes (50% pneumonia and 50% non-pneumonia) and different clustering algorithm has been used.

Results

In Table 34 results have been sorted in terms of less incorrectly clustered and then bigger likelihood. The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

The reason that the result from EM (real value) has been chosen are as follows:

1-The distribution of data in cluster 1(with good distribution) looks reasonable 31% of pneumonia(26 out of 83 patients) and 18% non pneumonia (16 out of 83 patients).(**A cluster which its majority is with pneumonia**).

2- **The percentage of incorrectly clusters instances are less than other algorithms**.

3-Comparing with other algorithms like EM (differences), which has a promising cluster, **it has the better log likelihood function** (since likelihood measures how likely clustering is, so the greater the log likelihood is, the better the clustering result is).

Algorithm	Cluster	Pnu (83 patients)	Non-pnu (83 patients)	percentage	Data	Log likelihood	**Incorrectly clustered instances**	Number of clusters achieved
EM	0	57	67	75%	Real	295	**44%**	2
	1	26	16	25%				
EM	0	67	76	86%	Difference	5.3	**44%**	2
	1	16	7	14%				
Make density based cluster	0	23	30	32%	Real	5.6	**46%**	2
	1	60	53	68%				
K-means	0	26	30	34%	Real	--	**47%**	
	1	57	53	66%				
K-means	0	26	30	34%	Difference	-----	**47%**	
	1	57	53	66%				
Make density based cluster	0	25	27	31%	Difference	268	**48%**	2
	1	58	56	69%				
Farthest first	0	79	82	97%	Real	---	**50%**	2
	1	4	1	3%				
Farthest first	0	82	83	99%	Difference	---------	**50%**	2
	1	1	0	1%				
cobweb	none				Real	---	**87%**	70
cobweb	none				Difference	---	**88%**	71
Hierarchical(Complete linkage, Euclidean)	1	83	83	**100%**	Difference	-------	**----**	---
Hierarchical (Complete linkage ,Euclidean)	0	3	9	7%	Real	----------	**--**	-------
	1	34	36	42%				
	3	46	38	51%				

Table 34: differences of applying different algorithm for clustering, using Weka and HCE. Blutgas. COHB

The knowledge representations of the promising cluster (cluster 1) of chosen algorithm (EM with Real value) from table 32 have shown in two respects.

1. Graph

Graph below (figure 18) shows the difference of cluster 0 with cluster 1 (as a promising one) from EM (Real value) As we see below only minor changes has been shown in the chart .For example the difference of Day stay 0 between cluster 0 and cluster 1 is only 0.07 is too small to be considered as a big difference between two clusters.

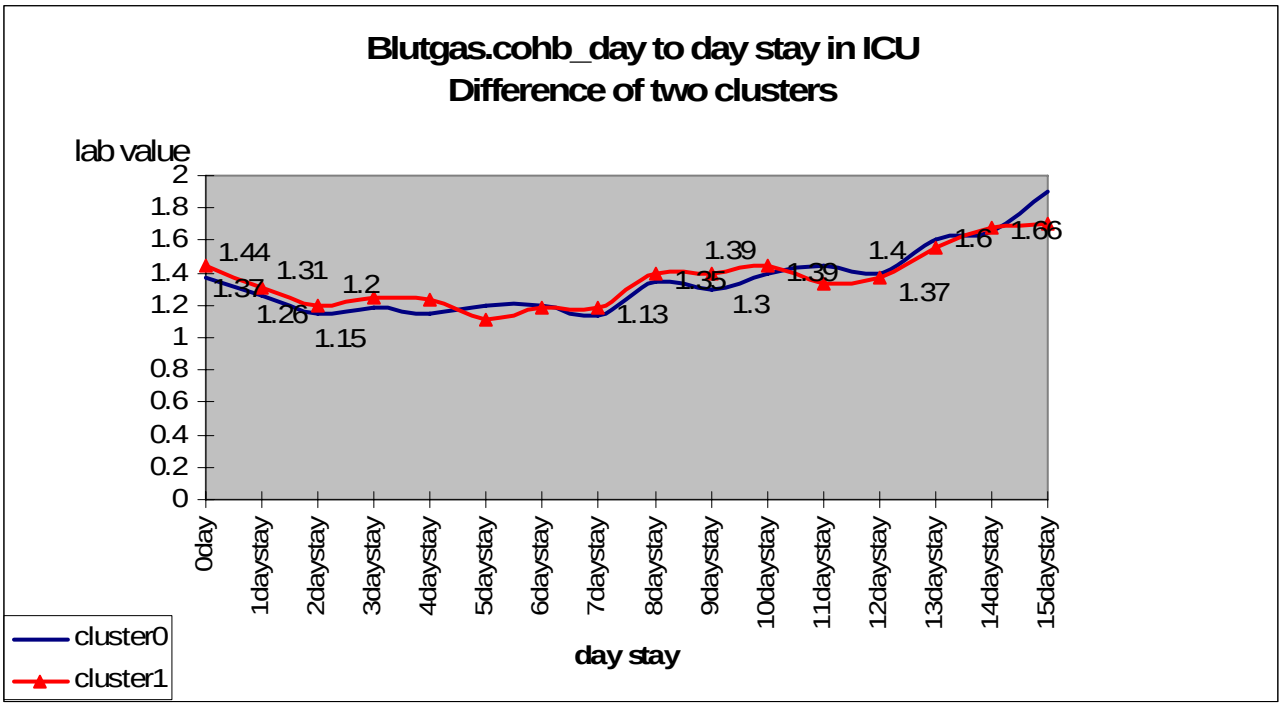

Figure 18: Lab value changes trend of Blutgas. COHB, difference of two clusters.

2. Table of difference between promising cluster(1) and other clusters

Since the same thing as we showed in Figure 18, it has been summarized in a table; it has not been shown here (for more information see **appendix B** table 16)

5.1.6. Conclusion of the results in Blutgas.COHB

By looking back to the previous chapters we got these results:

In part 5.1.1, we have assumed that two pneumonia and non pneumonia lab values are too close. This has been seen from the results from frequencies of lab values we have experienced to all clustering sections Finally we could not find any significant differences among cluster to make us to make a decision .In spite of promising results in part **5.1.5 but by studying the two clusters ,which shows the best cluster and the other clusters in figure 18 , presented that data are too close .Thus results are not reliable to make a decision even we have found promising results in joining pneumonia and non-pneumonia**

At this point we can say, when we can not find an attribute to separate two clusters from each others, we can not say we have found information to relay on, even when we get some good results in clustering.

What ever in Section "Some hypothesis and assumption from 5.1.1." has been assumed**, has been** experienced till here. First assumption is true; any promising cluster has had the less percentage of whole sample, as we see in table 32 in cluster1 from EM (Real).

Second assumption is not true because the differences (Δ) did not give the better results.

Third assumption is strongly true, the changes and differences is too small to find any pattern in promising cluster.

Now, the answers to the questions in **introduction**:

1-After clustering, any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why? Yes it has been achieved

by using EM and using Real value (value itself, not the difference among value of lab data day to day).

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

No, It has not been found any important attributes to make difference between promising cluster and the other clusters, because the difference among them are too small to be considered as a big difference between two clusters. It is concluded that finding a good distribution in a cluster (more pneumonia and less non-pneumonia) does not mean that there is definitely reliable ad interesting hidden pattern exist in it.

5.2 Blutgas_venös.COHB

(Blood Gas venous).HBO2.Normal arterial blood is bright red, whereas venous blood is slightly darker in color. HbO2 (Hyperbaric oxygen) is an Oxyhemoglobin fraction of the Blood gas Venus component .Its unit scale measure is %.

5.2.1 Preliminary analyses of Blutgas_venös.HBO2 of Frequency lab value

Goal

Here the goal is as the same as what was mentioned in part **"**4.1.1 Preliminary analyses of Blutgas.BE**"**goal. (To see complete tables and details in Appendix B.Tables 17, 18)
We will see big difference in the frequency order between pneumonia and non pneumonia, however the most frequent lab values are not too similar but in terms of percentage of distribution of lab values both pneumonia and non-pneumonia show the same distribution..

Observations

Figures 19, 20 we can conclude that both pneumonia and non-pneumonia have the same frequency values .In both pneumonia and non- pneumonia, the lab values are between 62 to 70 with Cumulative Percent more than 72 %(Cumulative percentage of lab value distribution), therefore in spite of having some changes in the most frequent lab value order pneumonia and non-pneumonias' lab values expect to have a close relation.

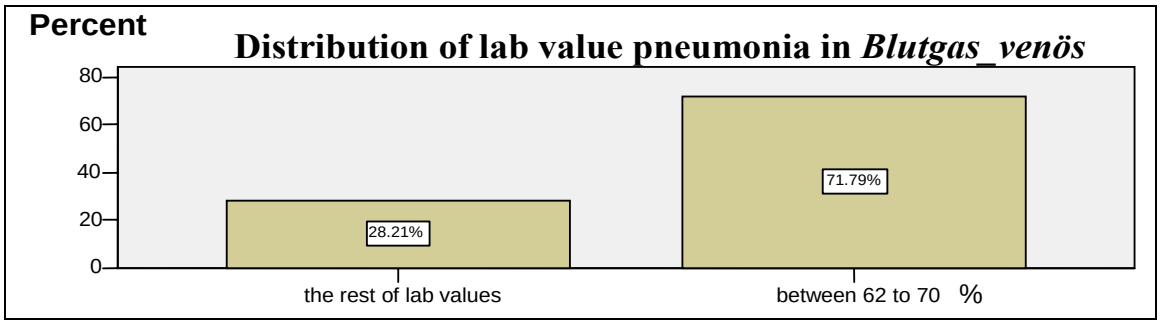

Figure 19: Distribution of lab value pneumonia in Blutgas_venös.

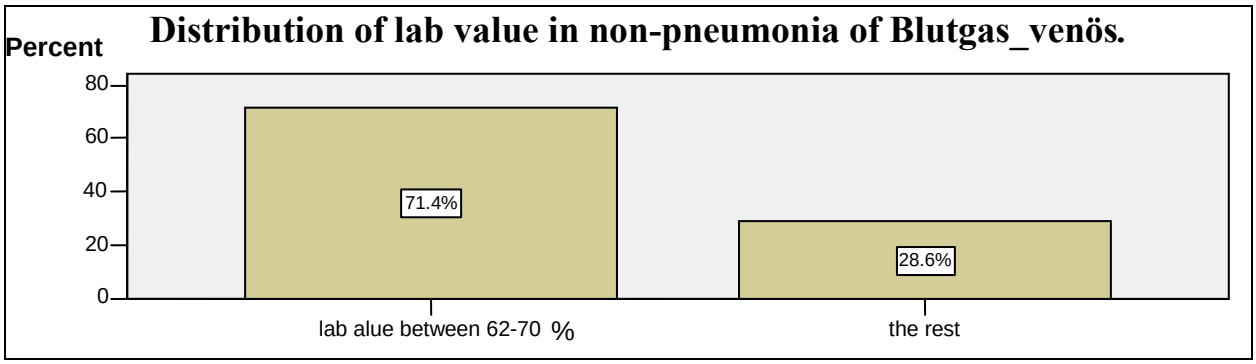

Figure 20: Distribution of lab value non-pneumonia in Blutgas_venös.

5.2.2 Preliminary analyses of Blutgas_venös.of out range of medical:

The medical range that was inserted for this lab values are as (92,100**)** (%).Over than 100 % never happen so we only have outrange of min (less than 92).

Technique

The same technique that was used in 4.1.2 is used here and following results have been achieved:

Results

In table 35, % number of patients that their labs values are out range of (92 % minimum of medical range) that is in the range 10 to 100% are 98% in pneumonia and 99% in non-pneumonia. Therefore, we can conclude that both pneumonia and no pneumonia cases have the tendency of going out range of minimum of medical range (92,100), So we first decided to replace the multi value in one day with minimum, although after no good result in clustering achieved comparing with when we used average. These results are showing again how much **lab values in pneumonia and non-pneumonia are close** and only minor differences that make a difference between promising cluster and other clusters.(for complete tables, **see appendix B table 19**).

Percentage of outrange (92%)min medical of lab value Profile B. Leukozyten	Number of patients in pneumonia (among 83 patients)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 856)	% number of patients in non-pneumonia
0-10%	1	1.2%	3	0.40%
10-100%	82	98.8%	853	99.6%

Table 35.Statistics value of Blutgas_venös out range (92%) in range (92,100) (Pneumonia and non-pneumonia patients)

5.2.3 Preparation stage of Blutgas_venös.HBO2 by separate clustering Pneumonia:

a) Pneumonia (average, real value)

It is clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) to see if there is any attributes that show a big difference among clusters?

The Attributes are used for clustering are: 74 including (age, sex, 0daystay, 1 day stay… 71daystay).

The results achieved by using EM algorithm are as follows:

Cluster numbers	Number of member	Log likelihood
0	4	-172.86
1	59	

Table 36: Clustering results of pneumonia in **Blutgas_venös.HBO2***.using real values.*

Cluster differences

Table 37 shows the difference between clusters achieved in this clustering by EM.

The interesting differences happened From Day stay 0 to Day stay 5 and also from Day stay 9 to Day stay 11 .There is a big difference between two clusters, since pneumonia and non-pneumonia lab values showed the same distribution and also cluster 0 has few members ,cluster 0 is usually considered as an outlier.

Cluster numbers	Name clusters	Number of members	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)								
0	A	4	Day stay 0	Day stay 1	Day stay 2	Day stay 3	Day stay 4	Day stay 5	Day stay 9	Day stay 10	Day stay 11
			45±0.7	73±0.6	81.1±0.5	80±0.6	82±0.5	74±0.7	55±0.6	48±0.9	54±0.87
1	B	59	65±0.09	68±0.07	68±0.07	69±0.08	69±0.7	68±0.05	73±0.5	72±0.03	74±0.05

Table 37: Summary of clustering all pneumonia of lab value Blutgas_venös.HBO2 with its clusters differences using average, real value.

c) Pneumonia (Δ, difference of data average)

The same approach that was done in section "a.5.2.3" has been followed here, except that, here the difference day to day of lab values instead of real values has been used. The results are in Table 38, cluster 0 contains outliers which has only two members .Since the results are rather similar to "a.5.2.3"in terms of difference between clusters, the tables of "difference between clusters" transferred to appendix b (**check table 20 in Appendix B**)

Attributes: 71 (age, sex, 0daystay, 1daystay… 68daystay)

Cluster numbers	Number of member	Log likelihood
0	2	36.03
1	61	

*Table 38: Clustering results of pneumonia in **Blutgas_venös.HBO2**.using **Δ**.*

5.2.4 Preparation stage of *Blutgas_venös.HBO2* by separate clustering Non-Pneumonia

a) Non-Pneumonia (average, real value)

The same approach that was done in section "a.5.2.3" has been followed here. The results are in Table 39.

Cluster numbers	Number of member	Log likelihood
0	90	**-48.8**
1	619	

*Table 39: Clustering results of non- pneumonia in **Blutgas_venös.HBO2**, using real value.*

Cluster differences

The difference between two clusters achieved in table 40 is as followed:

Cluster numbers	Name clusters	Number of member	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)				
0	**A1**	**90**	**Day stay 1**	**Day stay 2**	**Day stay 3**	**Day stay 4**	**Day stay 5**
			72±0.06	**74±0.02**	**71.±0.4**	**70.1±0.5**	**70±0.3**
1	**B1**	**619**	**71.4±1.08**	**70.9±0.6**	**69.9±0.18**	**67.5±0.05**	**67.2±o.7**

Table 40: Summary of clustering all non-pneumonia of lab value Blutgas_venös.HBO2 with its clusters differences using average, real value

b) Non-Pneumonia (Δ, difference of data average)

The same approach that was done in section "a.5.2.3" has been followed here. The results are in table 41, except that, here the difference day to day of lab values instead of real values has been used.

Cluster numbers	Number of member	Log likelihood
0	111	**-112.7**
1	598	

*Table 41: Clustering results of non- pneumonia in **Blutgas_venös.HBO2**, using **Δ**.*

Table of clusters differences can be found in **appendix B in table 21.**

5.2.5 Clustering Joining Pneumonia and Non-Pneumonia *Blutgas_venös.HBO2:*

Technique

The same technique that mentioned in 4.1.5 for Blutgas .BE. has been used here.

The sample includes (50% pneumonia and 50% non-pneumonia) and different clustering algorithm has been used.

Results

In Table 42 results have been sorted in terms of less incorrectly clustered and then bigger likelihood. The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

What has to be mentioned about Table ***42 is it*** shows the difference between applying different algorithms. **The reason that the result from EM using real value has been chosen** are as follows:

1-The distribution of data in cluster 0 is good, more pneumonia and less non-pneumonia.

2- **The percentage of incorrectly clusters instances are less than other algorithms**.

3-Comparing with other algorithms like Make density based cluster(real value) ,which have a promising cluster, **it has the better log likelihood function** .Likelihood measures how likely clustering is, so the greater the log likelihood is, the better the clustering is).

Algorithm	clusters	Pneumonia (63)	Non-pnu (63)	Percentage Of total sample	Data	Log likelihood	**Incorrectly clustered instances**	Number of clusters achieved
EM	0	3	1	2%	real	-44.9	**46%**	2
	1	60	62	94%				
Make density based cluster	0	0	1	1%	difference	-124	**49%**	2
	1	62	63	99%				
K-means	0	1	0	1%	difference	-----	**49%**	
	1	62	63	99%				
Farthest first	0	62	63	99%	real	---	**49%**	2
	1	1	0	1%				
Farthest first	0	62	63	99%	difference	-------	**49%**	2
	1	1	0	1%				
Make density based cluster	0	3	0	2%	real	-68	**50%**	2
	1	60	63	96%				
EM	0	1	1	2%	difference	-114	**50%**	3
	1	2	0	2%				
	2	49	50	79%				
	3	11	12	17%				
K-means	0	3	0	4%	real	---	**51%**	
	1	60	63	96%				
cobweb		---	------	---	real	--------	**96%**	126
cobweb		---	------	---	difference	--------	**96%**	126
Hierarchical (Complete linkage ,Euclidean)		63	63	100%	real	-----	---	-----
Hierarchical(Complete linkage, Euclidean)		63	63	100%	real	-----	---	-----

Table 42: difference of applying different algorithm for clustering using Weka and HCE. Blutgas_venös.HBO2

Discussion

In table 40 and also table 35 ,we can easily see the presence of outliers[22] ,first because ,an outlier is an unusually small or unusually large value in a data set [9] and is a small size of cluster [10],which can easily see in table 40,table 35 respectively.

The knowledge representations of the promising cluster (cluster 1) of chosen algorithm (EM with real value) from table 40 have shown below:

1. Graphs

Graph below (Figure 22) shows the difference among cluster 0and cluster 1.

Big difference have been observed at the first 6 days and Day stay stay 9 to 11and.All these observations show an abnormal distance from other values which confirm the outlier presentation.

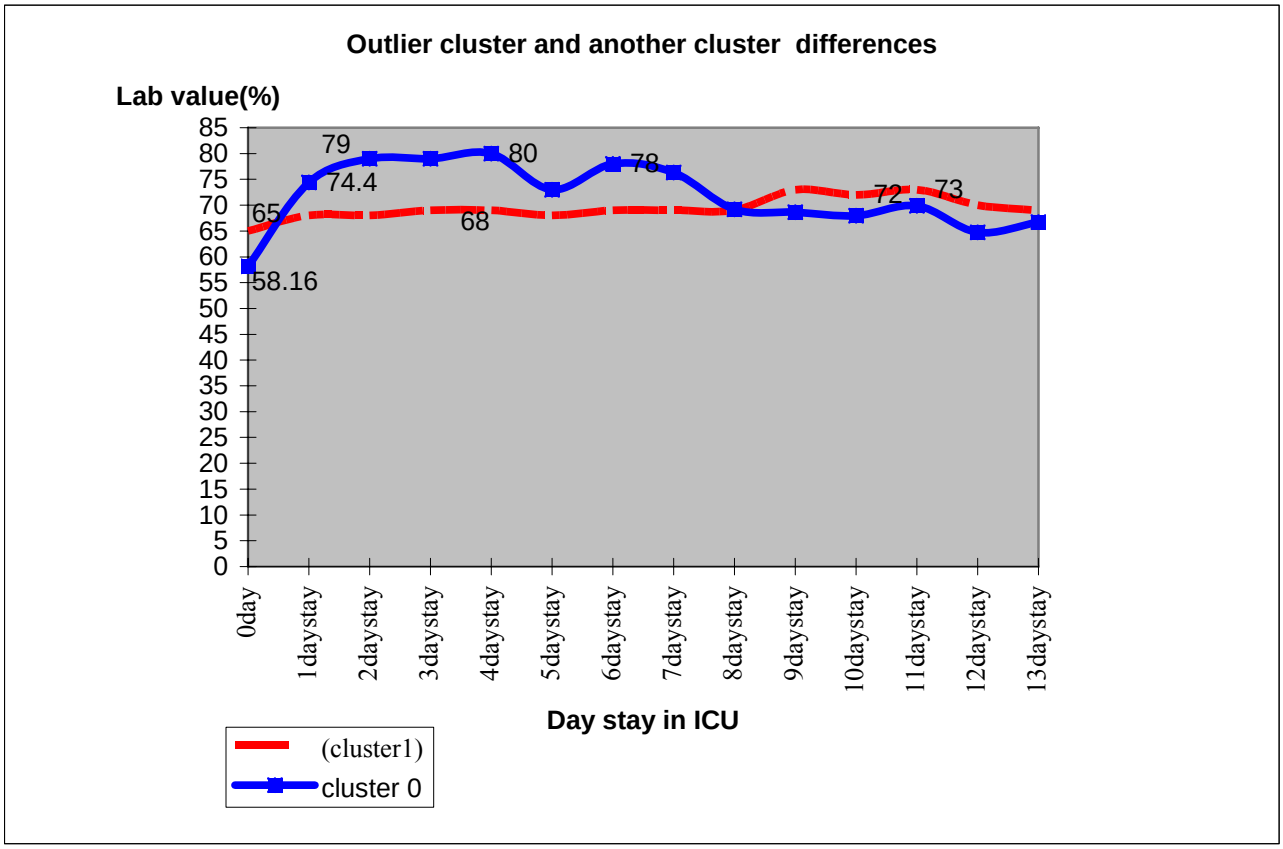

Figure 21: Cluster 0 (outliers) and cluster 1 Blutgas_venös.HBO2.

(See Appendix B Figure 1, outliers from non-pneumonia in cluster 0 and outliers from pneumonia in cluster 0)

Clusters differences

The table which shows the same thing that has been presented in Figure 21, so it has been skipped.

5.2.6. Conclusion of the results in Blutgas_venös.HBO2

In joining clusters, the outliers have been easily separated from samples. The reason is first a small size of cluster apperaed and they had a big distance with their neighborhood (another cluster).

The answers to the questions in **introduction**:

1-After clustering, any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why?

Yes it has been achieved by using EM.

2- If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

No although the distribution of clusters is reasonable but in a good distributed cluster (more pneumonia and less non-pneumonia) showed the big distance with another cluster and it had a small size of cluster.Therefore, they are considered as outlier not the promising .The attributes that showed these big differences are not enough to rely on.

Chapter 6

Lab values with no promising results

In this chapter we are dealing with lab values that their clustering results are not promising .Here I used only one lab value as an example and the rest has been listed as the similar case, because they do not give the robust and promising results that is relevant in point of clustering. In addition, since the primary analysis does not help in conclusion, a brief summary of this part has been mentioned, and also for the other part (clustering).

6.1. Profile.b.Thrombozyten

It is translated to Profile.b.Thrombocyte: A type of blood cell that helps to prevent bleeding by causing blood clots to form. Also called a platelet. Its measuring unit is Gpt/l (**Gigapartikel pro Liter**).

6.1.1. Frequency of lab value in pneumonia and non-pneumonia in Profile Thrombozyten:

According to the study of this lab value, it showed that they are so scatter and outspread, the most frequent lab values has been observed only in 9 cases out of 1054.Therefore, we are dealing with a wide range of values, no suitable cumulative percentage can help to separate the most frequent from others and also there is difference among most frequent lab values in pneumonia and non-pneumonia.

(Check the tables in Appendix B.tables

6.1.2 Preliminary analyses of Profile.b.Thrombozyten of out range of medical data:

The medical range that was inserted for this lab values are as (**150,300). (**Gpt/l).

Techniques

The same technique that has been used in 4.1.2 was used and following results have been achieved:

Results

In below tables 43, 44 do not show any more tendency of going out any out range neither out of minimum 150 nor out of 300 in pneumonia and non-pneumonia .The percentage of going out range of minimum in pneumonia and non-pneumonia are rather the same, it means both of them have the same tendency of going out of minimum (150).But in maximum out range in table 44, pneumonia shows more tendency of going out range of (300) than non-pneumonia. The only thing, we can conclude that is most of data are within the range of medical range (150,300) in non-pneumonia. (**Here has been just shown a portion of tables for more information, see appendix B table 24, 25**).

Percentage of outrange (150)min medical of lab value Profile.b.Thrombozyten	Number of patients in pneumonia (among 82)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 978)	% number of patients in non-pneumonia
0-10%	35	43%	440	45%
10-100%	47	57%	538	55%

*Table 43.Statistics value of Profile.b.Thrombozyten (150) in range (150,300**). (Pneumonia and non-pneumonia patients)*

Percentage of outrange 300 (max medical data) of lab value Profile.b.Thrombozyten	Number of patients in pneumonia (among 82)	% number of patients in pneumonia	Number of patients in non-pneumonia (978)	% number of patients in non-pneumonia
0-10%	10	12%	653	67%
10-100%	72	88%	328	34%

Table 44: Statistics value of Profile.b.Thrombozyten out range (300) in range (150,300). (pneumonia and non-pneumonia patients)

6.1.3 Preparation stage of *Profile.b.Thrombozyten* by separate clustering Pneumonia:

a) Pneumonia (average, real value)

The approach was started by clustering all pneumonia patients (the patients who stayed in ICU more than 3 days) .Among 680 pneumonia only 76 of them has this lab value and stay in ICU more than 3 days. The same approach that was done in section "a.5.2.3" has been followed here. The results are in Table 45.

cluster	Number of members	logliklihood
0	76	-260

Table 45: Clustering results of pneumonia in Profile.b.Thrombozyten, using real value.

As we see, no separation has been achieved.

b) Pneumonia (Δ, difference of data average)

The same approach as above in "6.1.3 a" but using difference of lab value day to day.

cluster	Number of members	logliklihood
0	76	-158.01

Table 46: Clustering results of pneumonia in Profile.b.Thrombozyten, using **Δ.**

In Table 46, as we see no separation here achieved

6.1.4 Preparation stage of *Profile.b.Thrombozyten* by separate clustering Non-Pneumonia

a) Non-Pneumonia (average, real value)

The same approach as above in "6.1.3 a"

The results are as follows:

cluster	Number of members	logliklihood
0	218	-344.27
1	750	

Table 47: Clustering results of non-pneumonia in Profile.b.Thrombozyten, using real value.

Clusters differences

Table 48 just show some attributes from two clusters as a sample.

Cluster numbers	**Number of member**	**Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)**				
		Day stay 1	**Day stay 2**	**Day stay 3**	**Day stay 4**	**Day stay 5**
0	218	**275±108**	**272±115**	**270±120**	**278±116**	**267±108**
1	750	**263±56**	**146±53**	**143±55**	**147±57**	**155±55**

Table 48: Summary of clustering all pneumonia of lab value Profile.b.Thrombozyten with its clusters differences using average, real value.

b) Non- Pneumonia (Δ, difference of data average)

The same approach as mentioned in section "6.1.4.a" but using day to day difference values.

cluster	Number of members	logliklihood
0	868	-116.21
1	100	

Table 49: Clustering results of non-pneumonia in Profile.b.Thrombozyten, using **Δ.**

(See cluster differences table in appendix B.table26)

6.1.5 Clustering Joining Pneumonia and Non-Pneumonia *Profile.b.Thrombozyten:*

Technique

The same technique that mentioned in 4.1.5 for Blutgas .BE. has been used here.

The sample includes (50% pneumonia and 50% non-pneumonia) and different clustering algorithm has been used.

Results

In Table 50 results have been sorted in terms of less incorrectly clustered and then bigger likelihood. The best algorithm has been highlighted. The percentage column shows how many percent of data in sample has been presented in each cluster.

In table 50, there is no good and promising separation achieved, in another word none of below algorithm could give a cluster with more pneumonia and less non-pneumonia.

What has to be mentioned about Table ***50 is,*** *it* shows the difference between applying different algorithms that none of them gave good results.

Algorithm	clusters	Pnu (76 patients)	Non-pnu (76 patients)	percentage	Data	Log likelihood	**Incorrectly clustered instances**	Number of clusters achieved
Make density based cluster	0	69	58	84%	difference	-233	**42%**	2
	1	7	18	16%				
Farthest first	0	25	29	34%	difference	----	**47%**	2
	1	51	47	49%				
Make density based cluster	0	74	76	99%	real	-250	**48%**	2
	1	2	0	1%				
K-means	0	24	27	34%	difference	-----	**48%**	2
	1	52	49	66%				
K-means	0	75	76	99%	real	----	**49%**	2
	1	1	0	1%				
Farthest first	0	75	76	99%	real	----	**49%**	-----
	1	1	0	1%				
EM		76	76	100%	real	-251	**50%**	1
EM	0	11	15	17%	difference	-178	**53%**	7
	1	0	1	1%				
	2	56	52	71%				
	3	2	0	1%				
	4	5	6	7%				
	5	1	0	1%				
	6	1	3	2%				
cobweb	none				real		**93%**	142
cobweb	none				difference	------	**96%**	146
Hierarchical (Complete linkage ,Euclidean)		76	76	100%	real	---	**-----**	1
Hierarchical(Complete linkage, Euclidean)		76	76	100%	difference	------	**-----**	1

Table 50: difference of applying different algorithm for clustering using Weka and HCE, in Profile.b.Thrombozyten.

6.1.6. Conclusion of the results in *Profile.b.Thrombozyten*

For this lab data has not found any promising cluster which has the more pneumonia and less non-pneumonia. For example in table 40, in EM (real value) no clustering separation has been seen ,or in EM (differences) there are separation but none of cluster have shown the good separation in terms of pneumonia and non pneumonia distribution. It was experienced in **6.1.3.also in** pneumonia; No separation has been achieved neither by real value nor with differences.

Now, answer to the **introduction** question for this lab value is:

1- After clustering any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why? No, none of them could make this separation.

2-If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses? No, since no promising cluster achieved ,so none of attributes can make the differences which is the consequences of what is achieved in preliminary analysis that has been concluded that data are so scatter and outspread.

The other lab values which gave the kind of non-promising results like Profile.b.Thrombozyten are as follows: **Blood gas** with components (**Gluc,Hbo2,lac,MetHB,pco2,PH,Po2**),**Hezenzyme,Troponin I ,Hezenzyme** (**h**eart enzymes) Leber Pankeras (Bilirubin (direct) [36],**Bilirubin(ges)** which is called Bilirubin total that measures the amount of bilirubin in a blood sample..[36]

Profile a.CRP: (see Appendix) is a test which measures the concentration in blood serum of a special type of protein produced in the liver.[37],**Harnstoff: (**Blood urea nitrogen (BUN)) measures the amount of urea nitrogen)**.[36].Kreatinin**:,the creatinine blood test to assess kidney function. [38]

Aptt :(Activated Partial Thromboplastin Time)**:** is used when someone has unexplained bleeding or clotting.[39]

Quick: A rapid and inexpensive blood test that measures levels of a hormone predicted the long-term health of patients with heart attack and chest pain. [40]

Chapter 7

Sex and age

According to experiments done for clustering, it showed that sex and age attributes did not affect the clustering results, it means by adding or deleting these attributes the result of clustering did not change. For example if EM algorithm used by using 74 attribute (sex,age,daystay0.....day stay 71) the results are the same as when EM algorithm used 72 attributes (daystay0.....day stay 71).In choosing samples from pneumonia and non-pneumonia ,it is also found out that when we just get sample.

For example only age range 60-79 which was most frequent in pneumonia in section "3.1.3.Sex and age frequencies in pneumonia" for choosing sample did not help to get the better result ,the result was most of the time worse than when we choose sample without considering specific age rang.

Conclusion and summary

This thesis was about clustering, and clustering analysis, in general working with lab data. In each chapter has been tried to answer 2 questions that was asked in introduction.These question were as follows:

1-After clustering any promising cluster (more pneumonia and less non-pneumonia) achieved if so, which algorithm gave the better results and why?

2-If promising results achieved, is there any important attribute(s) that have shown the significant role in separation of clusters? If not what is the reason? If so how these important attributes can mach with pneumonia days of diagnoses?

In Chapter 4 (Clustering lab values with pattern), the lab values have been clustered showed the promising results when they have been evaluated (classes to cluster evaluation) in terms of pneumonia and non-pneumonia distributions. Therefore at the end the promising clustering (more pneumonia and less non-pneumonia) presented by graph which shows the critical days .This contained attributes that have shown the significant changes comparing with other attributes .These differences have been presented in the tables that contains these attributes values (mean and standard deviation) that make differences between promising cluster and other clusters.

The pre-analyses, frequencies of lab values in each lab values, showed that how much lab values in pneumonia and non-pneumonia are close. Therefore, that has been concluded that we have to expect that the promising cluster less numbers of patients in joining pneumonia and non-pneumonia.(which was true, by observations).These results have been also observed more or less from medical outrange analysis .Although, this part has been done to find first if there is any significant difference among pneumonia and non-pneumonia in terms of going out of range of medical data and second to see the multiple lab values should be replaced by maximum or minimum or others .In general view ,this part showed that pneumonia and non-pneumonia have had the common behavior and again the same expectation in joining two sort of patients which promising cluster less numbers of patients. In final view, we would like to look at the achieved critical points (day stays) in terms of how much they were far from the day of pneumonia diagnosed .what has been resulted , these days have been set in the interval [-5,5] pneumonia happened. The answer to question 1 in this chapter is, Yes, EM because it gave the less incorrectly cluster instances and big log likelihood and also better

separation. The answer to question 2 is presenting the critical days and then matches the pneumonia patients in promising cluster with these critical days.

In chapter 5, (Clustering lab values (without pattern**).**This chapter has presented lab data that even they have shown some promising separation in joining pneumonia and non-pneumonia, but they have not shown any pattern. These lab data at the end either did not show any differences among the clusters, or they have shown they are outliers. In general, we have concluded that, promising cluster does not always mean the pattern has been found. The answer to question 1 in this chapter was yes it showed still EM gave the better result in terms of less incorrectly instance and bigger likelihood .The answer to question 2 is since the promising cluster did not showed big difference with other clusters or showed a big difference with small size comparing with other clusters (outlier) it can not be reliable results, so no knowledge achieved.

In chapter 6, that has been shown, the lab data which have neither promising cluster nor obviously any pattern. In most of them from pre-analysis, we have found that the lab values are so scatter in terms of frequency in pneumonia or non-pneumonia. (Not real frequent values found).This results was confirmed when in joining pneumonia and non-pneumonia, no promising clusters have been achieved. The answer to question 1 is No, since no promising separation achieved therefore no knowledge achieved .Thus the answer to question 2 is no promising cluster separation achieved so no critical days achieved.

In this thesis, it has been experienced that clustering approach goal is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be

independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. The main requirements that a clustering algorithm should satisfy are: [41]

Scalability; (applicable to huge databases), **Versatility**: be good at dealing with different types of attributes; **Ability to discover clusters with different shapes, Minimal input parameter:** require a minimum amount of domain knowledge, **Robust to noise, Insensitive to the data input order, Scaleable to high dimensionality.**

Historically, there is no single algorithm that can fully satisfy all the above requirements. It is important to understand the characteristics of each algorithm so the proper algorithm can be selected for the clustering problem.

The properties of EM which has showed the selected the best algorisms respect to result are: EM shows optimal results especially with noisy samples (robust in noise). Its lack of explanation requires additional analysis by a supervised learning model (has been done in chapter 4, 5).It has a reasonable Scalability (different amount sample in our data gave rather the same results in terms of finding the same hidden pattern).It had Versatility that was good at dealing with different types of attributes. It was insensitive to the data input order and it is easy to implement. Briefly speaking we can say, EM has a strong statistical basis. It is linear in database size. It provides a cluster membership probability per point. It can handle high dimensionality. It converges fast by given a good initialization.

Future work:

This research demands more efforts by using different methods which works on pneumonia, like regression and find out if a chosen patient is pneumonia or not, Sequential patterns to find e.g.the patients are more likely to get pneumonia.The work can be extended by working on diagnoises analyses for decision instead of lab values. [42]. For real medical data that are sequential, numerical, and ill-defined, by pre-processing methods and developed a rule discovery support system can obtained pattern combination rules in medical test results for pneumonia [43].

Figures / Tables

****

Appendix B

Blutgas arteriell.B.E

Blutgas arteriell.B.E Value range ((mmol/L)) pneumonia	Number of times the value observed among 1357 observations (Frequency)	Percent of Number of times the value observed among 1357 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1357 observations (Frequency lab value	Number of patients that the value have been observed (total 109 patients)	Percent Number of patients that the value have been observed (total 109 patients
2<=X<3	164	12.1	12.1	58	53
1<=x<2	148	10.9	23.0	59	54
3=<x<4	143	10.5	33.5	60	55
4<=x<5	133	9.8	43.3	57	52
0=<x<1	107	7.9	51.2	49	45
5<=x<6	98	7.2	58.4	52	48
-1<x<0	93	6.9	65.3	37	34
-1<=x<-2	86	6.3	71.6	31	28
6<=x<7	61	4.5	76.1	28	26
-2<=x<-3	51	3.8	79.9	23	21
7<=x<8	40	2.9	82.8	19	17
-3<=x<-4	35	2.6	85.4	16	15

Table1: A portion of tale of Frequencies lab value .pneumonia Blutgas.B.E

Blutgas arteriell.B.E Value range ((mmol/L)) Non-pneumonia	Number of times the value observed among 1357 observations (Frequency)	Percent of Number of times the value observed among 1357 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1357 observations (Frequency lab value	Number of patients that the value have been observed (total 109 patients)	Percent Number of patients that the value have been observed (total 109 patients
1<=x<2	1208	10.9	10.9	595	92.0
0=<x<1	1181	10.7	21.6	590	90.0
2<=X<3	1167	10.5	32.1	592	89.0
3=<x<4	1036	9.4	41.5	510	79.0
-1<x<0	907	8.2	49.7	473	69.0
4<=x<5	881	8.0	57.6	436	67.0
-1<=x<-2	758	6.8	64.5	420	58.0
5<=x<6	704	6.4	70.8	370	54.0
-2<=x<-3	495	4.5	75.3	290	38.0
6<=x<7	494	4.5	79.8	272	38.0
7<=x<8	348	3.1	82.9	186	27.0
-3<=x<-4	329	3.0	85.9	214	25.0
-4<=x<-5	222	2.0	90.3	153	17.0

*Table2: A portion of tale of Frequencies lab value_*B.E.*non-pneumonia*

Percentage of outrange (-2.3)min medical of lab value Blutgas.B.E	Number of patients in pneumonia (among 109)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 1313)	% number of patients in non-pneumonia
0-10%	82	75%	913	70%
10-20%	9	8%	116	9%
20-30%	9	8%	72	5%
30-40%	2	2%	65	5%
40%-50%	2	2%	47	4%
50-60%	0	0%	12	1%
60-70%	1	1%	19	1%
70-80%	0	0%	19	1%
80-90%	1	1%	7	1%
90-100%	3	3%	43	3%

Table 3.Statistics value of Blutgas out range (-2.3) in range (-2.3 2.3). (Pneumonia and non-pneumonia patients)

Percentage of outrange 2.3 (max medical data) of lab value Blutgas.B.E	Number of patients in pneumonia (among 109)	% number of patients in pneumonia	Number of patients in non-pneumonia (1313)	% number of patients in non-pneumonia
0-10%	20	18%	437	32%
10-20%	7	6%	82	6%
20-30%	8	7%	74	6%
30-40%	8	7%	83	6%
40%-50%	9	8%	104	8%
50-60%	7	6%	72	5%
60-70%	11	10%	61	5%
70-80%	8	7%	105	8%
80-90%	6	6%	40	3%
90-100%	25	23%	255	20%

Table4: Statistics value of Blutgas out range (2.3) in range (-2.3 2.3), (pneumonia and non-pneumonia patients)

Cluster numbers	Number of member	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)[19,21]					
0	107	Day stay 5	5Day stay 6	Day stay 15	Day stay 16	Day stay 17	Day stay 18
		2.1 ± 0.9	(2.6±0.4)	2..5±0.8	3.1±0.7	3.9±0.3	3.15±0.59
1	19	3.6±0.7	4.01±0.42)	2.9±0.7	4.5±0.19 shows there is decreasing from Daystay 15 to Day 16	2.6±0.3	4.09±0.59 (It showed that data has increased to range 4 from rang 2 until day 23)

Table 5: Summary of clustering all pneumonia of lab value Blutgas.B.E and its clusters differences using mean value of data.

Cluster numbers	Number of member	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)		Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)			
		Daystay5	Daystay6	Daystay7	Daystay14	Daystay15	Daystay16
0	97	2.0±0.4	2.3±0. 5	2.32±0.3	2.8±0.3	2.8±0.08	3.1±0.5
1	115	3.6±0.4	3.4±0. 7	3.00±0.8	3.5±0.1	3.7±0.3	4.6±0.44
2	847	2.06±0.9	2.06±0.3	3.00±0.8	2.6±0.3	2.7±0. 5	43.12±0
3	3	Not applicable(outliers)					

Table 6: Summary of clustering all non-pneumonia of lab value Blutgas.B.E and its clusters differences using mean value data.

Profile B.Leukozyten

Profile B. Leukozyten Value range ((Gpt/l)) pneumonia	Number of times the value observed among 1460 observations (Frequency)	Percent of Number of times the value observed among 1460 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1460 observations (Frequency lab value	Number of patients that the value have been observed (total 93 patients)	Percent Number of patients that the value have been observed (total 93 patients
8<=X<9	157	10.746	10.7	50	54%
9<=X<10	143	9.78	20.5	49	53%
10<=X<11	128	8.761	29.31	52	56%
11<=X<12	109	7.460	36.78	56	60%
13<=X<14	104	7.118	43.90	43	46%
7<=X<8	100	6.844	50.75	44	47%
12<=X<13	95	6.502	57.26	47	51%
14<=X<15	79	5.407	62.67	40	43%
15<=X<16	63	4.312	66.98	36	39%
6<=X<7	59	4.038	71.02	32	34%
17<=X<18	56	3.832	74.86	26	28%
16<=X<17	53	3.627	78.49	31	33%
18<=X<19	41	2.806	81.301	26	28%
5<=X<6	35	2.395	83.6	17	18%
19<=X<20	29	1.984	85.68	17	18%
20<=X<21	20	1.368	87.05	17	18%
21<=X<22	18	1.23	88.28	15	16%
4<=X<5	17	1.163	89.45	7	8%

Table 7: A portion of table of Frequencies lab value in Profile B. Leukozyten.

Profile B. Leukozyten Value range ((Gpt/l)) Non-pneumonia	Number of times the value observed among 11460 observations (Frequency)	Percent of Number of times the value observed among 11460 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 11460 observations (Frequency lab value	Number of patients that the value have been observed (total 995 patients)	Percent Number of patients that the value have been observed (total 995 patients
8<=X<9	1127	9.8	9.8	495	49%
9<=X<10	1086	9.5	19.3	487	49%
10<=X<11	1025	8.9	28.3	502	50%
7<=X<8	1019	8.9	37.1	461	46%
11<=X<12	944	8.2	45.4	480	48%
6<=X<7	801	7	52.4	355	35%
12<=X<13	760	6.6	59	427	43%
13<=X<14	682	6	65	375	37%
14<=X<15	541	4.7	69.7	312	31%
5<=X<6	453	4	73.6	236	24%
15<=X<16	436	3.8	77.4	280	28%
16<=X<17	380	3.3	80.8	243	24%
17<=X<18	332	2.9	83.6	212	21%
4<=X<5	251	2.2	85.8	118	12%
18<=X<19	239	2.1	87.9	154	15%
19<=X<20	211	1.8	89.8	146	15%
20<=X<21	168	1.5	91.2	122	12%
22<=X<23	132			91	9%

Table 8: A portion of tale of Frequencies lab value in Profile B. Leukozyten.non-pneumonia

Percentage of outrange (3.8)min medical of lab value Blutgas.HB.	Number of patients in pneumonia (among 93)	% number of patients in pneumonia	Number of patients in non-pneumonia (among (1001)	% number of patients in non-pneumonia
0-10%	89	95.7%	973	97.2%
10-20%	3	3.2%	14	1.4%
20-30%	0	0.0%	2	0.2%
30-40%	1	1.1%	3	0.3%
40%-50%	0	0.0%	1	0.1%
50-60%	0	0.0%	2	0.2%
60-70%	0	0.0%	1	0.1%
70-80%	0	0.0%	1	0.1%
80-90%	0	0	1	0%
90-100%	0	0	3	0%

Table7.1: Statistics value of Profile B.Leukozyten out range (3.8) in range (3.8 9.8),

(pneumonia and non-pneumonia patients)

Percentage of outrange	Number of patients in pneumonia (among 93)	% number of patients in pneumonia	Number of patients in	% number of patients in
(3.8)min medical			non-pneumonia (among 1001)	non-pneumonia
of lab value Blutgas.B.E				
0-10%	8	8.6%	158	15.8%
10-20%	3	3.2%	83	8.3%
20-30%	9	9.7%	39	3.9%
30-40%	6	6.5%	79	7.9%
40%-50%	6	6.5%	84	8.4%
50-60%	9	9.7%	80	8.0%
60-70%	8	8.6%	67	6.7%
70-80%	12	12.9%	92	9.2%
80-90%	8	8.6%	76	7.6%
90-100%	24	25.8%	243	24.3%

Table7.2: Statistics value of Profile B.Leukozyten range (9.8) in range (3.8 9.8),

(pneumonia and non-pneumonia patients

Blutgas.HB

Profile B. Leukozyten Value range (mmol/l) pneumonia	Number of times the value observed among 1325observations (Frequency)	Percent of Number of times the value observed among 1325observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1325 observations (Frequency lab value	Number of patients that the value have been observed (total 92 patients)	Percent Number of patients that the value have been observed (total 92 patients
(5<=X<6)	550	39.88	41.50	68	73.91%
(6<=X<7)	519	37.63	80.67	82	89.13%
(7<=X<8)	153	11.09	92.22	48	52.17%
(4<=X<5)	59	4.27	96.67	16	17.39%
(8<=X<9)	34	2.46	99.24	15	16.30%
(9<=X<10)	9	0.65	99.92	7	7.61%
(3<=X<4)	1	0.07	100	1	1.09%

Table 9: A whole table of Frequencies lab value of Blutgas.HB Pneumonia

Profile B. Leukozyten Value range ((mmol/l)) Non-pneumonia	Number of times the value observed among16513 observations (Frequency)	Percent of Number of times the value observed among 6513 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 6513observations (Frequency lab value	Number of patients that the value have been observed (total 813patients)	Percent Number of patients that the value have been observed (total 813 patients
(5<=X<6)	3169	48.64	48.66	567	69.74%
(6<=X<7)	1915	29.39	78.06	533	65.56%
(7<=X<8)	598	9.17	87.24	250	30.75%
(4<=X<5)	520	7.98	95.22	213	26.20%
(8<=X<9)	135	2.07	97.30	81	9.96%
(0<=X<1)	76	1.16	98.46	34	4.18%
(9<=X<10)	42	0.64	99.11	25	3.08%
(3<=X<4)	34	0.52	99.63	25	3.08%
(2<=X<3)	14	0.21	99.85	10	1.23%
(10<=X<11)	6	0.09	99.94	5	0.62%
(1<=X<1)	3	0.04	99.98	3	0.37%
(11<=X<4)	1	0.01	100.00	1	0.12%

Table 10: A whole table of Frequencies lab value of Blutgas.HB Pneumonia

Percentage of outrange (3.8)min medical	Number of patients in pneumonia (among 92)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 813)	% number of patients in non-pneumonia
10-20%	0	0%	0	0.0%
20-30%	0	0%	0	0.0%
30-40%	0	0%	2	0.2%
40%-50%	1	1%	3	0.4%
50-60%	0	0%	2	0.2%
60-70%	0	0%	3	0.4%
70-80%	0	0%	7	0.9%
80-90%	3	3%	10	1.2%
90-100%	88	96%	770	94.7%

*Table 11***:** *Statistics value of Blutgas .HB outrange (8.6) in range (8.6,12), (pneumonia and non-pneumonia patients*

Blutgas. COHB

Blutgas. COHB Value range (%) pneumonia	Number of times the value observed among 1082 observations (Frequency)	Percent of Number of times the value observed among 1082 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1082 observations (Frequency lab value	Number of patients that the value have been observed (total 100 patients)	Percent Number of patients that the value have been observed (total 100 patients
1=<x<2	614	56.74	56.74	76	76.00%
2=<x<3	272	25.13	81.88	53	53.00%
0=<x<1	184	17	98.89	30	30.00%
3=<x<4	12	1.109	100	8	8.00%

*Table 12***:** *A whole table of Frequencies lab value of Blutgas. COHB Pneumonia.*

Blutgas. COHB Value range (%) pneumonia	Number of times the value observed among 9140 observations (Frequency)	Percent of Number of times the value observed among 9140 observations (Frequency	Cumulative Percent of Number of times the value observed among 9140 observations (Frequency lab	Number of patients that the value have been observed (total 1142 patients)	Percent Number of patients that the value have been observed (total 1142 patients

		lab value	value		
1=<x<2	5562	60.85	60.85	997	87.30%
0=<x<1	2145	23.46	84.32	485	42.50%
2=<x<3	1295	14.16	98.4	345	30.20%
3=<x<4	97	1.06	99.5	44	3.90%
4=<x<5	15	0.16	99.71	10	0.90%
-1<x<0	9	0.09	99.81	4	0.40%
5=<x<6	4	0.04	99.85	3	0.30%
6=<x<7	3	0.03	99.89	3	0.30%
15=<x<16	3	0.03	99.92	3	0.30%
-38<x<=-39	1	0.01	99.93	1	0.10%
-18<x<=-19	1	0.01	99.94	1	0.10%
-14<x<=-15	1	0.01	99.95	1	0.10%
0<x<=-1	1	0.01	99.96	1	0.10%
7=<x<8	1	0.01	99.97	1	0.10%
11=<x<12	1	0.01	99.98	1	0.10%
12=<x<13	1	0.01	100	1	0.10%

Table 13: A whole table of Frequencies lab value of Blutgas. COHB non-Pneumonia.

Cluster numbers	**Number of member**	**Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)**			
0	19	**Day stay 10**	**Day stay 11**	**Day stay 12**	**Day stay 13**
		3.1±0.3	**2.99±0.3**	**3.4±0.01**	**3.37±0.007**
1	69	**3.3±0.02**	**3.01±0.02**	**3.55±0.02**	**3.4±0.02**

Table 14:: The difference between two clusters achieved using EM Blutgas COHB in pneumonia using real value

Cluster numbers	Number of member	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD			
0	104	Day stay 11–Day stay 10	Day stay 12-Day stay 11	Day stay 13-Day stay 12	Day stay 14-Day stay 13
		(0.0001 ±0.3	-0.0004±0.2	0.02 ±0.19	-0.03±0.2
1	748	0.01±0.07	0.008±0.06	0.018±0.07	0.07 ±0.05)
2	25	0.02±0.6	0.03±1.2	0.13±0.8	0.3 ±1.9
3	10	0.04 ±0.2	-0.04 ±0.2	-0.15 ±0.2	0.1198±0.3

Table 15: The difference between two clusters achieved using EM Blutgas COHB in non-pneumonia using Δ

Cluster numbers	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD (real value day to day)				Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD) (Differences day to day)			
0	Day stay 1	Day stay 2-Day stay 1	Day stay 3-Day stay 2	Day stay 4-Day stay 3	Day stay 1	Day stay 2	Day stay 3	Day stay 4
	(-0.11 ±0.5	-0.07 ±0.4	-0.015 ±0.2	-0.03 ± 0.2	1.26 ±0.06	1.15 ±0.02	1.19 ±0.21	1.15 ± 0.65
1	(-0.12 ±0.4	-0.1 ±0.2	-0.015 ±0.1	0.04 ±0.1	1.31 ±0.08	1.2 ±0.04	1.25 ±0.5	1.23 ±0.6

Table 16: differences found in cluster 1 and cluster 0 using EM (differences and real value) of blutgas.cohb.

Blutgas_venös.HBO2

Blutgas_venös.HBO2 Value range (%) pneumonia	Number of times the value observed among 957 observations (Frequency)	Percent of Number of times the value observed among 957 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 957 observations (Frequency lab value	Number of patients that the value have been observed (total 83 patients)	Percent Number of patients that the value have been observed (total 83 patients
68<=X<69	51	5.3	5.3	37	45.00%
72<=X<73	49	5.1	10.4	26	31.00%
70<= X <71	43	4.5	14.9	28	34.00%
71<= X <72	43	4.5	19.4	23	28.00%
73<= X <74	42	4.4	23.8	22	27.00%
74<= X <75	42	4.4	28.2	20	24.00%
78<= X <79	42	4.4	32.6	17	20.00%
77<= X <78	41	4.3	36.9	19	23.00%
75<= X <76	38	4.0	40.9	21	25.00%
76<= X <77	34	3.6	44.4	23	28.00%
65<= X <66	33	3.4	47.9	24	29.00%
67<= X <68	33	3.4	51.3	22	27.00%
69<= X <70	32	3.3	54.6	24	29.00%
66<= X <67	30	3.1	57.8	19	23.00%
79<= X <80	30	3.1	60.9	18	22.00%
64<= X <65	29	3.0	63.9	22	27.00%
80<= X <81	26	2.7	66.7	18	22.00%
63<= X <64	25	2.6	69.3	19	23.00%
62<= X <63	24	2.5	71.8	17	20.00%

Table 17: *A portion of table of Frequencies lab value Blutgas_venös.HBO2, pneumonia.*

Blutgas_venös.HBO2 Value range (%) Non-pneumonia	Number of times the value observed among 8093 observations (Frequency)	Percent of Number of times the value observed among 8093 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 856 observations (Frequency lab value	Number of patients that the value have been observed (total 856 patients)	Percent Number of patients that the value have been observed (total 856 patients
69<= X <70	373	4.6	4.6	254	30%
70<= X <71	372	4.6	9.2	257	30%
71<= X <72	362	4.5	13.7	242	28%
68<= X <69	361	4.5	18.1	240	28%
72<= X <73	360	4.4	22.6	235	27%
75<= X <76	351	4.3	26.9	220	26%
74<= X <75	347	4.3	31.2	233	27%
73<= X <74	333	4.1	35.3	221	26%
76<= X <77	323	4.0	39.3	207	24%
66<= X <67	321	4.0	43.3	225	26%
77<= X <78	309	3.8	47.1	204	24%
67<= X <68	307	3.8	50.9	217	25%
64<= X <65	276	3.4	54.3	204	24%
78<= X <79	256	3.2	57.5	163	19%
65<= X <66	255	3.2	60.6	187	22%
79<= X <80	233	2.9	63.5	159	19%
63<= X <64	232	2.9	66.4	186	22%
62<= X <63	213	2.6	69.0	168	20%
80<= X <81	194	2.4	71.4	131	15%

Table 18: A portion of table of Frequencies lab value Blutgas_venös.HBO2 Non_Pneumonia

Percentage of outrange (92%)min medical of lab value Blutgas.B.E	Number of patients in pneumonia (among 83)	% number of patients in pneumonia	Number of patients in non-pneumonia (among 856)	% number of patients in non-pneumonia
0-10%	1	1.2%	3	0.4%
10-20%	0	0.0%	0	0.0%
20-30%	0	0.0%	0	0.0%
30-40%	0	0.0%	0	0.0%
40%-50%	0	0.0%	3	0.4%
50-60%	0	0.0%	0	0.0%
60-70%	2	2.4%	9	1.1%
70-80%	1	1.2%	10	1.2%
80-90%	3	3.6%	18	2.1%
90-100%	76	91.6%	813	95.0%

Table 19: *.Statistics value of Blutgas out range (-2.3) in range (-2.3 2.3). (Pneumonia and non-pneumonia patients*

Cluster numbers	Number of members	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)									
0	2	Day stay 1-Daystay 0	Day stay 2-Day stay 1	Day stay 3-Daystay 2	Day stay 4-day stay 3	Day stay 5-daystay 4	Day stay 6-Day stay 5	Day stay 8-Day stay 7	Day stay 9-Day stay 8	Day stay 10-Day stay 9	Day stay 11- Day stay 10
		30±0.7	-2±0.7	-5±0.6	8±0.6	-13±0.5	14±0.7	-13±o.7	-12±0.3	-6±0.2	6±0.3
1	2	1.2±0.9	0.2±0.4	1.07±7	-0.4±0.08	1.6±0.3	0.5±5	-0.4±0.99	2.8±0.3	-0.4±0.4	0.9±0.3

Table 20: Summary of clustering all pneumonia of lab value Blutgas_venös.HBO2 with its clusters differences using Δ, difference of data average day to day

Table of differences can be found in appendix B in table 21.

Cluster numbers	Number of member	Comparison of Cluster s by Δ mean (daystayK+1 –Day stay K) ± STD				
0	111	Day stay 1-Day stay0	Day stay 2-Day stay1	Day stay 3-Day stay 2	Day stay 4-Day stay 3	Day stay 5-Day stay 4
		0.5± 0.03	0.9±0.5	71.±0.4	-1.02±0.2	-1.6±0.6
1	598	-0.2±0.02	-1.12±0.4	69.9±0.18	-0.8±0.4	-0.96±0.08

Table 21: Summary of clustering all non-pneumonia of lab value Blutgas_venös.HBO2 with its clusters differences using Δ

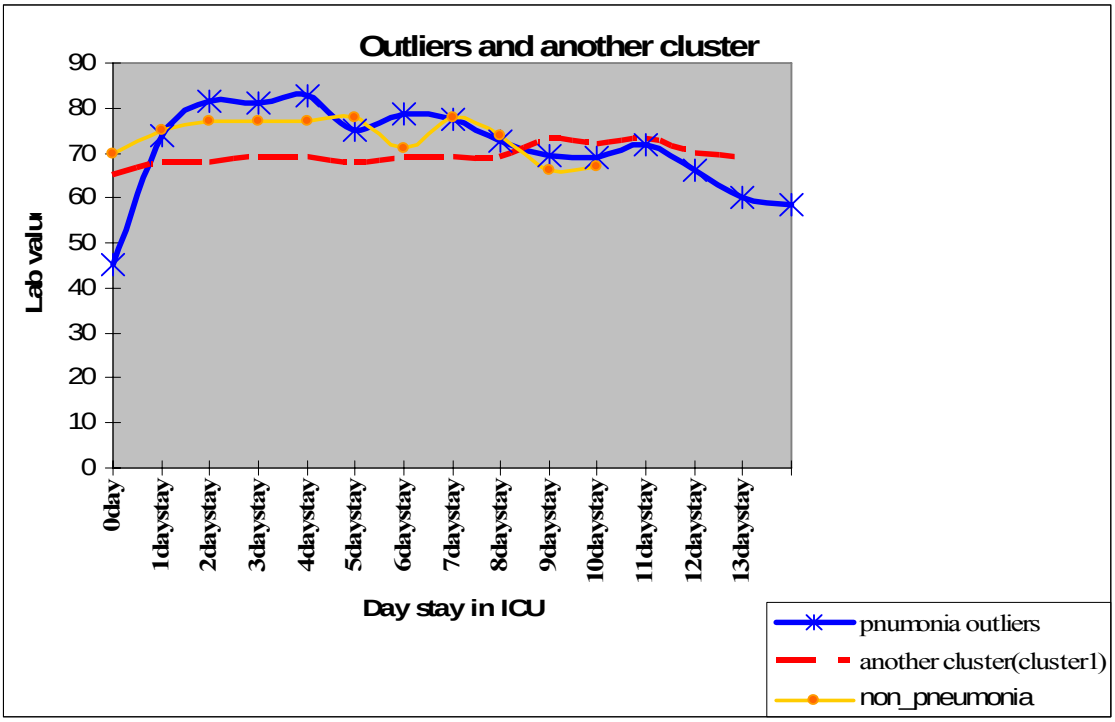

Figure 1: presentation of the outlier cluster and another clusters in Blutgas_venös.HBO2.

Profile.b.Thrombozyten

Profile.b.Thrombozyten Value range ((Gpt/l)) pneumonia	Number of times the value observed among 1054 observations (Frequency)	Percent of Number of times the value observed among 1054observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 1054 observations (Frequency lab value	Number of patients that the value have been observed (total 82patients)	Percent Number of patients that the value have been observed (total 82 patients
224	9	0.9	1.7	7	9%
139	9	0.9	0.9	6	7%
216	8	0.8	3.2	7	9%
186	8	0.8	2.5	8	10%
295	7	0.7	3.9	5	6%
310	6	0.6	9.6	5	6%
234	6	0.6	9.0	5	6%
229	6	0.6	8.4	6	7%
220	6	0.6	7.9	5	6%
215	6	0.6	7.3	6	7%
201	6	0.6	6.7	6	7%
154	6	0.6	6.2	6	7%
148	6	0.6	5.6	6	7%
126	6	0.6	5.0	6	7%
113	6	0.6	4.5	6	7%
389	5	0.5	21.4	5	6%
372	5	0.5	21.0	4	5%
371	5	0.5	20.5	2	2%

Table 22: A portion of tale of Frequencies lab value in pneumonia Profile.b.Thrombozyten.

Profile.b.Thrombozyten Value range ((Gpt/l)) Non-pneumonia	Number of times the value observed among 11451 observations (Frequency)	Percent of Number of times the value observed among 11451 observations (Frequency lab value	Cumulative Percent of Number of times the value observed among 11451 observations (Frequency lab value	Number of patients that the value have been observed (total 1001 patients)	Percent Number of patients that the value have been observed (total 1001 patients
146	66	0.576369	0.576368876	66	0.07
144	58	0.506506	1.082874858	56	0.06
156	55	0.480307	1.563182255	55	0.05
114	54	0.471575	2.03475679	54	0.05
179	53	0.462842	2.497598463	53	0.05
96	52	0.454109	2.951707274	52	0.05
122	51	0.445376	3.397083224	51	0.05
124	50	0.436643	3.833726312	49	0.05
142	50	0.436643	4.2703694	50	0.05
151	50	0.436643	4.707012488	50	0.05
192	50	0.436643	5.143655576	50	0.05
120	49	0.42791	5.571565802	49	0.05
140	49	0.42791	5.999476028	48	0.05
160	49	0.42791	6.427386254	48	0.05
164	49	0.42791	6.855296481	49	0.05
167	49	0.42791	7.283206707	49	0.05
182	49	0.42791	7.711116933	49	0.05
187	48	0.419177	8.130294297	46	0.05
121	47	0.410445	8.5407388	46	0.05
136	47	0.410445	8.951183303	47	0.05
138	47	0.410445	9.361627805	47	0.05
141	47	0.410445	9.772072308	47	0.05
112	46	0.401712	10.17378395	46	0.05
130	46	0.401712	10.57549559	46	0.05
150	46	0.401712	10.97720723	45	0.04
158	46	0.401712	11.37891887	46	0.05
165	46	0.401712	11.78063051	45	0.04
184	46	0.401712	12.18234215	46	0.05
204	46	0.401712	12.58405379	46	0.05

Table 23: A portion of tale of Frequencies lab value in non-pneumonia Profile.b.Thrombozyten.

Profile.b.Thrombozyten

Percentage of outrange 150 (max medical data) of lab value *Profile.b.Thrombozyten*	Number of patients in pneumonia (among 82)	% number of patients in pneumonia	Number of patients in non-pneumonia (978)	% number of patients in non-pneumonia
0-10%	35	43%	440	45%
10-20%	4	5%	72	7%
20-30%	2	2%	46	5%
30-40%	6	7%	64	7%
40%-50%	7	9%	44	4%
50-60%	4	5%	37	4%
60-70%	5	6%	38	4%
70-80%	5	6%	60	6%
80-90%	4	5%	45	5%
90-100%	10	12%	132	13%

*Table 24.Statistics value of Profile.b.Thrombozyten (150) in range (150,300***).** *(Pneumonia and non-pneumonia patients)*

Percentage of outrange 300 (max medical data) of lab value *Profile.b.Thrombozyten*	Number of patients in pneumonia (among 82)	% number of patients in pneumonia	Number of patients in non-pneumonia (978)	% number of patients in non-pneumonia
0-10%	10	12%	653	67%
10-20%	5	6%	66	7%
20-30%	3	4%	27	3%
30-40%	8	10%	50	5%
40%-50%	10	12%	41	4%
50-60%	7	9%	27	3%
60-70%	9	11%	36	4%
70-80%	12	15%	20	2%
80-90%	6	7%	12	1%
90-100%	12	15%	49	5%

Table25: Statistics value of Profile.b.Thrombozyten out range (300) in range (150,300). (pneumonia and non-pneumonia patients

Cluster numbers	Number of member	Differences between Cluster 0 and cluster 1 by comparisons of average (value average ± STD)				
		Day stay 1-Day stay0	Day stay 2-day stay1	Day stay 3-Daystya2	Day stay 4-Daystay 3	Day stay 5-Day stay 4
0	218	-10±35	-13±32	-3±32	5.7±29	9.7±23
1	750	-12±33	14±32	-4±28	8.5±22	4.1±23

Table 26: Summary of clustering all non-pneumonia of lab value Profile.b.Thrombozyten with its clusters differences using Δ

References:

[1]: Bruce Carlson, 2001 Nidus Information Services Inc,*Well-Connected reports June 2001*, a board of physicians, including faculty at Harvard Medical School and assachusetts General Hospital Cynthia Chevins http://www.reutershealth.com/wellconnected/doc64.html".

[2]: The EM Algorithm for Unsupervised Clustering: http://grb.mnsu.edu/grbts/doc/manual/Expectation_Maximization_EM.html)

[3]: Fast clustering in SQL using the EM algorithm**:** Carlos Ordonez and Paul Cereghini, Dallas, Texas, United States, Year of Publication: 2000, ISSN: 0163-5808

[4]: COBWEB Algorithm for Incremental Clustering: http://grb.mnsu.edu/grbts/doc/manual/COBWEB.html.

[5]: Feature Article: Data Mining: An AI Perspective: Data Mining: An AI Perspective Xindong Wu[1], *Senior Member, IEEE* . Ming-Syan Chen, Jiawei Han, and Philip Yu, Data Mining: An Overview from a Database Perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8 (1996), 6: 866-883.

[6]: B. Mac Queen (1967),]: *A Tutorial on Clustering Algorithms,* Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297,refered to, Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides",http://www-.cs.cmu.edu/~awm/tutorials/kmeans.html ,Brian T. Luke: "K-Means Clustering",http://fconyx.ncifcrf.gov/~lukeb/kmeans.html ,Tariq Rashid,"Clustering",http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clusteri ng_initial_report/node11.html ,Hans-Joachim Mucha and Hizir Sofyan: "Nonhierarchical Clustering".

[7]: Sameer Agarwal_ Ravi,"*Structured Importance Sampling of Environment Maps" (2004)* Ramamurthy† Serge Belongie_ Henrik Wann Jensen__University of California, San Diego†Columbia University.

[8]: Steve Simons: http://www.childrensmercy.org/stats/.

[9]: Yanxia Zhang, Ali Luo, Yongheng ZHAO, National Astronomical observatories, Chinese Academy of Sciences(2004), *Outlier detection in astronomical data,* refered to E. Hung, D. W. Cheung, *Parallel Algorithm for Mining Outliers in Large Database", in Distributed and Parallel Database*, Kluwer Academic Publisher, 12, 2002.

[10]:Ben-Gal I., Kluwer Academic Publishers (2005), Maimon O, and Rockach L. (Editors.). *Data Mining and Knowledge Discovery Handbook: a Complete Guide for Practitioners and Researchers*, chapter" *Outlier detection*"

[11]: Tatiana Semenova1, Markus Hegland2, Warwick Graco3, and Graham Williams41 (2004),CSL, *Conceptual Mining of Large Administrative Health Data,*RSISE, Australian National University, School of Mathematical Sciences, Australian National University Australian Health Insurance Commission, Mathematical and Information Sciences, CSIRO.

[12]: Semih UTKU, (July, 2004), *MINING ASSOCIATION RULE ALGORITHMS IN LARGE DATABASES*, A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Dokuz Eylül University.

[13]: Arun George Eapen, 2004, *Application of Data mining in Medical Applications* , thesis presented to the University of Waterloo, in Systems Design Engineering, Waterloo, Ontario, Canada.

[14]: Usama Fayyad*, Gregory Piatetsky-Shapiro, and Padhraic Smyth,* **(1997),** *From Data Mining to Knowledge Discovery in Databases,* Ai Magazine, Conferences/National/1997/aaai97.html

[15]: World Health Organization: http://www.who.int/classifications/icd/en/.

[16] Lokesh S. Shrestha, (March 25, 2004), *Machine learning with Weka,* Columbia university.

[17]: Sergiu Chelcea1, Alzennyr Da Silva1&2, Yves Lechevallier2, Doru Tanasa1, Brigitte Trousse, AxIS, INRIA Sophia-Antipolis(2004), Route des Lucioles, B.P. 93,Sophia Antipolis Cedex, France, AxIS, INRIA Rocquencourt,Domaine de Voluceau, Rocquencourt, 78153 Le Chesnay Cedex, France, *Pre-Processing and Clustering Complex Data in ECommerce ,Domain* ,http://www-sop.inria.fr/axis/.

[18]: Simona Korenjak- _ Cerne, Vladimir Batagelj**,**(July 15-19 2002),*Symbolic data analysis approach to clustering large datasets,*University of Ljubljana,Slovenia IFCS 2002, Cracow, Poland.

[19]: Thomas Gärtner, Shaomin Wu, and Peter A. Flach,*Data Mining on the Sisyphus Dataset: Evaluation and Integration of Results* , Department of Computer Science, University of Bristol, Woodland Road, Bristol BS8 1UB, U.K. Knowledge discovery team, AiS, GMD, Schloß Birlinghoven, D-53754 Sankt Augustin, Germany.

[20]: Berkhin P. & Software A., (2002), Survey of Clustering Data Mining Techniques.

[21]: Statistics Canada, *Variance and standard deviation,* (downloaded: 2006-12-07) http://www.statcan.ca/english/edu/power/ch12/variance.htm.

[22]: Ville Hautam¨aki, Ismo K¨arkk¨ainen and Pasi Fr¨anti, *Outlier Detection Using k-Nearest Neighbour Graph,*University of Joensuu, Department of Computer Science,Joensuu, Finland.

[23]:Martin Bland(10 August 2006), *Applied Biostatistics Mean and Standard Deviation,* University of York ,Heslington ,UK.

[24]: Steffen Bickel and Tobias Scheffer,(2004), *Multi-View Clustering*,Humboldt-Universit¨at zu Berlin,Department of Computer Science,Unter den Linden 6, 10099 Berlin, Germany.

[25]: Frank Dellaert,(February 2002)*The Expectation Maximization Algorithm,* College of Computing, Georgia Institute of Technology, Technical Report number GIT-GVU-02-20.

[26]: List of technical and popular medical terms: German (b) http://users.ugent.be/~rvdstich/eugloss/DE/lijstb.html ,(downloaded: 2006-04-07).

[27]: Lloyd-Williams, M. "*Case studies in the data mining approach to health information analysis*", *Knowledge Discovery and Data Mining (1998/434),* IEEE Colloquium on,8 May1998, 1996 Page(s): 1/1 -1/4.

[28]: Holmes G., Donkin A. and Witten I.H. (1994) "*WEKA: Machine Learning Workbench" Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia..

[29]: Wang, Y. and Wong, A.K.C.; from association to classification: inference-using weight of evidence, *IEEE Transactions on Knowledge and data engineering,* Volume: 15, Issue: 3, May-June 2003 Pages: 764 – 767

[30]: W. Wang, J. Yang, and R. Muntz. STING: *A statistical information grid approach to spatial data mining*. In *Proceedings* of 23rd International Conference on Very Large Data Bases*,* pages 186-195, Athens, Greece, 1997.

[31]: Glymour, C.; Scheines, R.; Spirtes, P.; Kelly, K. 1987.*Discovering Causal Structure.* New York: Academic.Guyon, O.; Matic, N.; and Vapnik, N. 1996. Discovering Informative Patterns and Data Cleaning. In *Advances in Knowledge Discovery and Data Mining,*eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 181–204. Menlo Park, Calif.: AAAI Press.

[32]: Jain, A. K., and Dubes, R. C. 1988. *Algorithms for Clustering Data.* Englewood Cliffs, N.J.: Prentice-Hall.

[33]: Zembowicz, R., and Zytkow, J. 1996. From Contingency Tables to Various Forms of Knowledge in Databases. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, 329–351. Menlo Park, Calif.: AAAI Press.

[34]: Anli K.Jain, Richard C.Dubes.1948 Algorithms for clustering Data, Prentice Hall Englewood Cliffs, New Jersey Michigan state University

[35]:Professor C O'Callaghan FRCH,DM, *Pneumonia in children,* Department of Infection, Immunity and Inflammation, University of Leicester and Department of Biological Sciences, University of Warwick, (downloaded 2006-01-01) http://www.action.org.uk/research_projects/grant/285/.

[36]: WebMD Medical Reference from Healthwise, (Last Updated: June 30, 2004)
http://www.webmd.com/digestive-disorders/Bilirubin-15434,
http://www.webmd.com/a-to-z-guides/Blood-Urea-Nitrogen

[37]: Carol & Richard Eustice,(2007), *What Is C-Reactive Protein (CRP)?* About, Inc., A part of The New York Times Company. http://arthritis.about.com/cs/diagnostic/a/crp.htm.

[38]: American Association for Clinical Chemistry(2001-2007) A public resource on clinical lab testing from the laboratory professional who do the testing, http://www.labtestsonline.org/understanding/analytes/creatinine/test.html

[39]: American Association for Clinical Chemistry(2001-2007) A public resource on clinical lab testing from the laboratory professional who do the testing,
http://www.labtestsonline.org/understanding/analytes/aptt/test.html

[40]: Torbjŕn Omland, M.D., Ph.D.; Anita Persson, M.Sc.; Leong Ng, M.D., Ph.D.; Russel O'Brien, M.D.; Thomas Karlsson, M.Sc.; Johan Herlitz, M.D., Ph.D.; and Marianne Hartford, M.D., Ph.D. Quick, cheap blood test predicts chance of surviving heart attack, **(**November 2002**),** American Heart Association,
http://www.scienceblog.com/community

[41]:Osmar R. Za¨_ane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang, *On Data Clustering Analysis: Scalability,Constraints and Validation,*University of Alberta, Edmonton, Alberta, Canada.

[42]:Joseph L. Hellerstein and Sheng Ma IBM T.J.*MINING EVENT DATA FOR ACTIONABLE PATTERNS,* Watson Research Center Hawthorne, New York

[43]: Miho Ohsaki1, Yoshinori Sato2, Hideto Yokoi3, and Takahira Yamaguchi1
A Rule Discovery Support System for Sequential Medical Data In the Case Study Of a Chronic Hepatitis Dataset **,** Faculty of Information, Shizuoka University
Japan, Graduate School of Information, Shizuoka University, Department of Medical Informatics, Chiba University Hospital.

[44]: Tom Mitchell 1997 "*Machine Learning*", McGraw Hill.