Journal of King Saud University – Computer and Information Sciences (2014) **26**, 347–354

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Software reuse in a paralysis dataset based on categorical clustering and the Pearson distribution

M. Bhanu Sridhar [a], Y. Srinivas [b], M.H.M. Krishna Prasad [c,*]

[a] *Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam, India*
[b] *Department of Information Technology, GITAM University, Visakhapatnam, India*
[c] *Department of Computer Science and Engineering, JNTU, Kakinada, India*

Received 21 April 2013; revised 30 September 2013; accepted 14 December 2013
Available online 20 December 2013

KEYWORDS

Reuse;
Paralysis;
Expert systems;
Pearson family of distributions;
Metrics;
Categorical clustering

Abstract Software reuse is the process of building software applications that make use of formerly developed software components. In this paper, we explain the benefits that can be obtained from using statistical procedures for prescribing medicines, especially in rural areas, which have limited resources available on hand. It should be noted that although the expert systems were successful in research, they never dominated the market when actual patient treatment was considered. The proposed methodology is compared with the categorical clustering technique. The Fenton and Melton Coupling Metric is considered for the evaluation of the statistic model. The reliability of this methodology is also considered.

Contents

[*] Corresponding author. Tel.: +91 9490458740.
E-mail addresses: sridharbhanu@gmail.com (M. Bhanu Sridhar), ysrinivasit@rediffmail.com (Y. Srinivas), krishnaprasad.mhm@gmail.com (M.H.M. Krishna Prasad).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

1319-1578 © 2013 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jksuci.2013.12.003

1. Introduction

Software reuse has long been identified as a key methodology for the improvement of software features. This process saves time, money and energy and also improves the quality of the new software. The success of the reuse process depends on many criteria, such as, the reusability of the existing software (Godin et al., 1995), the availability of efficient algorithms to obtain the maximum reusability, and so on. Software reuse has long been attracting researchers who recognize its worth in reducing the cost and time for building a new project. Software reuse has long been successful in the research field but not in industry. Furthermore, reuse in the medical field has not received the required attention due to the early introduction of expert systems. In this paper, expert systems and their approach to utilize medical data are explained and compared with the idea of utilizing statistical methodologies. Furthermore, the pros and cons of both angles are discussed to vindicate faith in the long-proven statistical methods. Throughout this paper, some of the statistical methods and their applications in the medical field are discussed and compared with expert systems. The article culminates with a conclusion that opens the doors for a new idea and its application, especially in rural areas, to save the precious lives of any ailing patients.

2. Discussion of expert systems

Expert systems (ESs) are a type of applied artificial intelligence (AI), which is a term that was coined in 1960s. The basic idea of these systems is expertise, which is the collection of task-specific knowledge, is transferred from a human to a computer (Liao, 2004). This knowledge is then stored in the system and becomes convenient for users to apply in specific cases. ESs are broadly divided into different categories, such as rule-based systems, knowledge-based systems, neural networks, fuzzy ESs, case-based reasoning (CBR), intelligent agent (IA) systems, modeling, and ontologies. Together with their applications, certain ESs for medical data are available in the literature, such as MYCIN, DENDRAL, INTERNIST, and CASNET.

ESs are used for specific purposes, such as investigating the side effects of a specific drug or predicting the disease of a patient from the answers to some queries that are posed. However, these methodologies have never thrown light upon the data that are available online in specific situations in which a patient from a remote area must be treated with an appropriate life-saving drug in stages, such as heart stroke, paralysis and other diseases, which result in either deformities or loss of life. In such a case, questions are not answered; the results from the tests conducted on the patient previously or presently are considered. Some work of this type has been reported (Bhanu Sridhar et al., 2012) where an attempt is made to investigate the reusability criteria on persons living in remote areas. Further work is also continued by the authors (BhanuSridhar et al., 2012), where they have highlighted the need for statistical models for clustering instead of the existing methodologies such as K-means algorithm.

It can be carefully noted that ESs are used mainly for predicting the disease of a patient or the side-effects of medicines but not for prescribing medicines for a patient. Additionally, ESs have never been used with confidence in the medical field (http://en.wikipedia.org/wiki/Expert_system, 2013). The knowledge collection that we use for the purpose of expert systems is usually error-prone due to human errors. Additionally, expert systems use computational engines incapable of reasoning and hence lead to incorrect conclusions. The logic that is used can be based upon facts that will surely change after a period of time, further degrading the level of the expert systems in the medical field.

Time is now appropriate to bring out new approaches by using more proven and stable methodologies, to be more helpful to the patients. Proceeding from this perspective angle, the ever-safe and sound statistical models fit easily into the required frame. They have been used for many decades and it is time now to apply them in medical field applications to prescribe medicines for the patients, with more confidence and assurance. Solid models such as the Gaussian Mixture Model and the Pearson Family of Equations come into the picture immediately and the model to be applied in the current situation must be decided upon. The model that is the most viable currently is to be discussed and decided upon.

3. Dataset

The dataset that is considered here is that of patients who have the symptoms of paralysis. Paralysis is a loss of muscle function due to sudden damage caused to the spinal cord. A study conducted by the Christopher and Reeve Foundation (2012) suggests that ~1 in 50 people have been diagnosed with paralysis. There exists very large number of reasons for the occurrence of paralysis and perhaps many more reasons that are yet to be discovered. Paralysis can result from diseases that either involve changes in the makeup of nervous or muscular tissue or are the result of metabolic disturbances that interfere with the function of nerves or muscles (Medline Net, xxxx). In this context, the application of Software Reuse is apparently more likely to enable the current patients to be able to use the substantial amount of result-oriented and promising medicine that depends on previous data, and thus, its reuse.

Paralysis occurs when something goes wrong with the way that messages pass between the brain and muscles. Paralysis can be complete or partial (Medline Plus, 2012). It can occur on one or both sides of the body. It can also occur in only one area, or it can be widespread. Paralysis of the lower half of the body, including both legs, is called paraplegia. Paralysis of the arms and legs is quadriplegia.

Usually, paralysis is due to strokes or injuries such as spinal cord injury or a broken neck. Other causes of paralysis include nerve diseases such as amyotrophic lateral sclerosis, autoimmune diseases such as Guillain–Barre syndrome and Bell's palsy, which affect the muscles in the face. The general symptoms of the disease are alcoholism, altered smell, numbness, weakness in the ocular muscles (eye), decreased reflexes, decreased sensations, balance problems, altered pulse rate, weakness in tongue, confusion, disorganized thinking, disability and high BP.

Software Reuse in the medical field is as continuously useful as reusing the medicines themselves. The data collected for a patient with a problem such as paralysis, which can be reused in the context of other patients to deduce whether the concerned patients are close to having a certain disease, are diagnosed with a disease or do not have the disease. Similarly, the medicines used previously under similar conditions will be very convenient when a patient with a similar condition is detected. Reuse is counted to be vital in the medical field since previous information is very convenient in deducing a patient's current health position and saving a precious life (Patil and Kumaraswamy, 2009). The database of the paralysis patients from the archives (Datasets from Machine Learn, 2012) is considered to apply the stated methods and comparisons. This circumstance can be more specific to situations in which a patient from a remote area must be treated with an appropriate life-saving drug in stages of urgency.

4. The Pearson family of probability distributions

In the medical realm, accuracy and efficiency are given utmost importance because assumptions and non-realistic conclusions could lead to disasters. It is better to classify the patients into different clusters, in which each cluster forms a set of patients who have similar symptoms and, hence, have the same prescriptions for medicines. Traditional clustering algorithms such as the Gaussian Mixture Model and K-means fail to cluster the medical data accurately because these methods consider all the attributes of the data during the process. In the case of medical data, which contains very large amounts of data, certain attributes could be irrelevant and, therefore, need not be considered during the clustering process. Consideration of these irrelevant attributes could falsify the clustering results (Paul, 2010). Moreover, another disadvantage of traditional clustering algorithms is that they are very sensitive to outliers (McLachlan and Peel, 2000). This disadvantage of outliers degrades the pattern recognition (Sun et al., 2010; Peel and McLachlan, 2000).

To overcome these disadvantages, the Pearson family is chosen because the robustness of the estimation of the Pearson family is very high (Sun et al., 2010). In the literature, the studies reported in BhanuSridhar et al. (2012) have been confined to the usage of the Pearson Type-I family. However, since the medical data are to be processed effectively and efficiently, the distributions that have heavy tails are much needed.

The main advantage of using the Type-VII distribution, which has been used here, is that the data that fall on the outliers can also be considered by assigning non-zero probabilities to the data that are away from the main cluster. The Pearson Type-VII also has heavy tails which help to cluster the data more accurately (Sun et al., 2010). The parameters of the distribution are updated by using the EM algorithm and the updated equations are considered for the development of the model. The dataset that is considered for the model has been explained in section-3, and the proposed methodology is discussed in section-5, using the coupling metric of Fenton and Melton. The dataset is translated to be in binary form, and categorical clustering is also applied to the dataset; the results derived for the model are presented in section-6. The comparative conclusions that are drawn from both clustering techniques are presented in the last section.

Karl Pearson proposed a family of distributions, which are popularly known as Pearson's distributions; they can be generated from the solution of the differential equation given below:

$$\frac{df(x)}{dx} = \left\{ \frac{b + x}{a_0 + a_1 x + a_2 x^2} \right\} f(x) \tag{1}$$

The random variable X denotes the gray level intensity of the echocardiograph speckle and $f(x)$ represents the probability density function (pdf). The b, a_0, a_1, and a_2 are parameters of the distribution. These parameters are determined in terms of the first four central moments (μ_i for i = 1, 2,..., 4) of the underlying empirical distribution. By using the method of moments, we have

$$b = -a_1 = \frac{0.5\mu_3(3\mu_2^2 + \mu_4)}{9\mu_2^3 - 5\mu_2\mu_4 + 6\mu_3^2} \tag{2}$$

$$a_0 = \frac{0.5\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{9\mu_2^3 - 5\mu_2\mu_4 + 6\mu_3^2} \tag{3}$$

$$a_2 = \frac{0.5(6\mu_3^2 - 2\mu_2\mu_4 + 3\mu_3^2)}{9\mu_2^3 - 5\mu_2\mu_4 + 6\mu_3^2} \tag{4}$$

Pearson identified a selection parameter κ that is expressible in terms of the first four moments. Defining Skewness ($S_k = \mu_3/\mu_2^{3/2}$) and Kurtosis ($K_u = \mu_4/\mu_2^{2}$), we find

$$\kappa = \frac{S_k^2(K_u + 3)^2}{4(4K_u - 3S_k^2)(2K_u - 3S_k^2 - 6)} \tag{5}$$

The three major types of Pearson densities, i.e. Type I, Type IV and Type VI are defined for $\kappa < 0$, $0 < \kappa < 1$ and $\kappa > 1$, respectively. We consider the pdf of the type-I distribution ($k < 0$) which is given by

$$f_1(x) = A_0 \left(1 + \frac{x - m_0}{c_1}\right)^{g_1} \left(1 - \frac{x - m_0}{c_2}\right)^{g_1}, \quad - c_1 + m_0 < x < c_2 + m_0 \tag{6}$$

where

$$A_0 = \frac{g_1^{g_1} g_2^{g_2} \Gamma(g_1 + g_2 + 2)}{(c_1 + c_2)(g_1 + g_2)^{g_1+g_2} \Gamma(g_1 + 1)\Gamma(g_2 + 1)} \tag{7}$$

$$g_{2,1} = 0.5h - 1 \pm sign(\mu_3)(0.5h(h + 2)) \times \frac{S_k}{\sqrt{S_k^2(h + 2)^2 + 16h + 16}} \tag{8}$$

$$c_1 = \left(\frac{g_1}{g_2}\right) \frac{0.5\sqrt{\mu_2\{S_k^2(h + 2)^2 + 16h + 16\}}}{1 + \frac{g_1}{g_2}} \tag{9}$$

$$c_2 = 0.5\sqrt{\mu_2\{S_k^2(h + 2)^2 + 16h + 16\}} - c_1 \tag{10}$$

$$h = \frac{6K_u - 6S_k^2 - 6}{6 + 3S_k^2 - 2K_u} \tag{11}$$

$$m_0 = \mu_1 - 0.5\frac{\mu_3(h+2)}{\mu_2(h-2)} \tag{12}$$

Here $b_1 = 0$ and b_0, b_2 have the same signs. The probability density function of Type – VII is given by

$$f(x) = c(1+\frac{x^2}{a^2})^{-m}, -\infty < x < \infty \tag{13}$$

where $m > \frac{1}{2}$

With

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{14}$$

$$\text{and } c\int_{-\infty}^{\infty}\left(1+\frac{x^2}{a^2}\right)^{-m} dx = 1 \tag{15}$$

Let

$$z = \left(1+\frac{x^2}{a^2}\right)^{-1} \tag{16}$$

$$=> x^2 = a^2\frac{(1-z)}{z}$$

$$=> x = a(1-z)^{\frac{1}{2}}z^{\frac{-1}{2}} \tag{17}$$

From Eq. (16),

$$dz = -\left(1+\frac{x^2}{a^2}\right)^{-2}\left(\frac{2x}{a^2}\right)dx)$$

$$=> dx = \frac{a}{z}z^{\frac{-3}{2}}(1-z)^{\frac{-1}{2}}dz \tag{18}$$

The final values of μ_1^I and μ_2^I are given below in Eqs. (19) and (20). Note that the detailed derivation is also presented.

Consider using Eqs. (16) and (18) in Eq. (15) and from the properties of definite integrals, we have

$$= c.\int_0^1 \frac{a}{2}z^m.z^{\frac{-3}{2}}(1-z)^{\frac{-1}{2}}dz = 1 \tag{19}$$

$$= c.\frac{a}{2}\int_0^1 z^m.z^{\frac{-3}{2}}(1-z)^{\frac{-1}{2}}dz = 1 \tag{20}$$

Rewriting Eq. (20) we have

$$\frac{a}{2}.c\int_0^1 (z)^{m-\frac{1}{2}-1}(1-z)^{\frac{1}{2}-1}dz = 1$$

$$=> \frac{a}{2}.c.\beta\left(m-\frac{1}{2},\frac{1}{2}\right) = 1 \tag{21}$$

$$=> c = \frac{2}{a.\beta\left(m-\frac{1}{2},\frac{1}{2}\right)} \tag{22}$$

where $a > 0$.

Using Eqs. (18) in (13) we have

$$f(x) = \frac{2}{a.\beta\left(m-\frac{1}{2},\frac{1}{2}\right)}\left(1+\frac{x^2}{a^2}\right)^{-m}, -\infty < x < \infty \tag{23}$$

The first two moments of the Pearson distribution are

$$\mu_1' = \int_{-\infty}^{\infty} x.f(x)dx \tag{24}$$

From Eq. (13)

$$= c\int_{-\infty}^{\infty} x\left(1+\frac{x^2}{a^2}\right)^{-m} dx \tag{25}$$

$$= c\int_0^1 z^m\left(\frac{a^2}{2}\right)z^{-2}dz$$

$$= \frac{a^2}{2}c\int_0^1 z^{m-2}dz \tag{26}$$

$$Put\ z = \left(1+\frac{x^2}{a^2}\right)^{-1}$$

$$= \frac{a^2}{2}c\left(\frac{z^{m-1}}{m-1}\right)$$

$$= \frac{a^2}{2}c\left(\frac{1}{m-1}\right) \tag{27}$$

From Eq. (22) we get

$$= \frac{a^2}{2}\frac{2}{a.\beta\left(m-\frac{1}{2},\frac{1}{2}\right)}.\frac{1}{m-1} \tag{28}$$

$$\mu_1' = \frac{a}{\beta\left(m-\frac{1}{2},\frac{1}{2}\right)}.\frac{1}{m-1} \tag{29}$$

By the definition of moments,

$$\mu_2' = \int_{-\infty}^{\infty} x^2 f(x)dx$$

From Eq. (12)

$$\mu_2' = \int_{-\infty}^{\infty} x^2 f(x)dx$$

$$= c.a^2\int_{-\infty}^{\infty}\left(1+\frac{x^2}{a^2}-1\right)\left(1+\frac{x^2}{a^2}\right)^{-m} dx \tag{30}$$

$$= c.a^2\int_{-\infty}^{\infty}\left(1+\frac{x^2}{a^2}\right)^{-m+1} - c.a^2\int_{-\infty}^{\infty}\left(1+\frac{x^2}{a^2}\right)^{-m} dx$$

Using Eq. (15)

$$= c.a^2\int_{-\infty}^{\infty}\left(1+\frac{x^2}{a^2}\right)^{-m+1} dx - a^2 \tag{31}$$

Using Eq. (18) we have

$$= c.a^2\int_0^1 z^{m-1}\left(\frac{a}{2}\right)z^{\frac{-3}{2}}(1-z)^{\frac{1}{2}-1}dz - a^2 \tag{32}$$

$$= c.\frac{a^3}{2}\int_0^1 z^{\left(m-\frac{3}{2}\right)^{-1}}(1-z)^{\frac{1}{2}-1}dz - a^2$$

$$= c.\frac{a^3}{2}\beta\left(m-\frac{3}{2},\frac{1}{2}\right) - a^2 \tag{33}$$

from the properties of $\beta(m,n)$.Using Eq. (22)

$$= \frac{a^3}{2}\left[\frac{2}{a.\beta\left(m-\frac{1}{2},\frac{1}{2}\right)}\right]\beta\left(m-\frac{3}{2},\frac{1}{2}\right)-a^2 \tag{34}$$

$$\mu_2' = a^2\frac{\beta\left(m-\frac{3}{2},\frac{1}{2}\right)}{\beta\left(m-\frac{3}{2}+1,\frac{1}{2}\right)}-a^2$$

$$= a^2\frac{\left(m-\frac{3}{2}+\frac{1}{2}+1\right)}{m-\frac{3}{2}}-a^2$$

$$= a^2\left(\frac{2m}{2m-3}\right)-a^2$$

$$=> \mu_2' = \frac{3a^2}{2m-3} \tag{35}$$

Using these values in equation-(6), the clustering process is performed.

The main advantage of using Pearson's Family of Distributions is its degree of freedom which controls the robustness. Among the different types of distribution that are available in this family of equations, Type-VII is much preferred due to the specific advantage of possessing heavy tails that help the user to cluster the data more accurately. These values are utilized for the clustering of the dataset against the patients' diseases and the results are tabulated in the tables that follow.

5. Categorical clustering

Cluster analysis plays an important role in the fields of data mining, statistics and informatics. In this methodology, the objects are divided into clusters or groups, and each cluster contains similar objects. Categorical data are concentrated on in this study, which emphasizes its relatedness with medical field. Usually, categorical data consists of categorical variables that are used for experiential data whose value is one of a fixed number of assumed categories, i.e., grouped data (http://en.wikipedia, 2013). In these cases, data are produced in cross-tabulation or matrices.

Categorical variables are characterized by values that are categories (Řezanková, 2009). The main types of these variables are dichotomous variables (binary) and multi-categorical variables, which are further divided into nominal, ordinal and quantitative. It should be noted that arithmetic operations can be applied only for quantitative variables and can calculate the distance between the objects to bring out a categorical clustering classification.

In the considered medical data about paralysis, binary data are considered in which 11 symptoms of paralysis are quoted for 50 patients. If a symptom is found, then it is represented by a value of 1 in the matrix; otherwise, a 0 is used. Apparently, objects are to be characterized by binary variables in the process of creating the matrix, and subsequently a hierarchical cluster analysis is performed. The dataset is shown in Table 1 below.

The 11 symptoms of paralysis that are considered here are Alcoholism (S1), ALS (S2), Genetic Problems (S3), Botulism (S4), Cauda Equina (S5), Cerebral Palsy (S6), Coxsackie Virus (S7), Nerve Problems (S8), Guillian–Barre Syndrome (S9), High Blood Potassium (S10) and Stroke (S11).

A two-way frequency table is used for the objects $\boldsymbol{x_i}$ $\boldsymbol{a}$ $\boldsymbol{x_j}$ as shown below in Table 2:

For symmetric variables, where 1 and 0 hold equal priority, Sokal and Michener's simple matching coefficient is usually used and, hence, is applied here. It is given by

$$S_{SM} = \frac{a+d}{a+b+c+d}$$

The coefficient is an algorithm that is considered to measure the distances between the components for their classification into clusters, by utilizing the k-Medoids methodology (Nascimento et al., 2012). In this approach, instead of taking

Table 1 Dataset.

Patient ID (↓)	Alcoholism (S1)	ALS (S2)	Genetic Problems (S3)	Botulism (S4)	Cauda Equina (S5)	Cerebral Palsy (S6)	Coxsackie Virus (S7)	Nerve Problems (S8)	Guillian–Barre Syndrome (S9)	High Blood Potassium (S10)	Stroke (S11)
P1	0	1	0	1	0	1	1	0	0	0	0
P2	0	0	0	1	1	1	1	0	0	0	0
P3	0	0	0	1	0	0	0	0	1	0	0
P4	0	0	0	0	0	0	0	0	0	0	0
P5	0	1	0	1	0	1	1	0	0	0	0
P6	0	0	1	1	1	1	1	1	1	1	1
P7	0	0	0	0	0	0	0	0	0	1	0
P8	0	0	0	1	1	1	1	0	0	0	0
P9	0	0	0	0	0	1	0	0	0	1	0
P10	0	0	1	0	1	0	1	0	1	0	1
P11	0	1	0	0	0	1	1	1	1	0	1
P12	0	0	1	0	0	0	0	0	1	1	0
P13	0	0	0	1	1	1	1	0	0	0	0
P14	0	1	0	0	1	1	1	1	0	0	0
P15	0	1	0	0	1	1	0	0	0	0	1
P16	0	1	0	1	0	1	1	0	0	0	0
P17	0	1	0	0	1	1	0	0	0	0	1
P18	0	0	0	1	0	0	1	1	1	1	0
P19	0	1	0	0	1	1	0	0	0	0	1
P20	0	1	0	1	0	1	1	0	0	0	0

Table 2 Frequency table.

Category of object x_i	Categories of object x_j	
	1	0
1	*a*	*b*
0	*c*	*d*

In this notation, a – number of attributes with both i and j present; b – number of attributes with only i present; c – number of attributes with only j present; d – number of attributes with both i and j absent.

the mean for the values, as in K-Means, a representative component is chosen to represent the whole cluster. Through a Java program, the distance between each pair of components is calculated by utilizing the matching coefficient cited above, and the output is placed in an MS-Excel sheet, which can be seen in Table 3. Observe that P1, P2…P20 are the patients, and the distances can be seen in the Table 3.

The outcome of the program suggests five clusters with numbers that match the types of Paralysis – Paraplegic, Paraplegia, Quadraplegic, Quadraplegia and Cerebral Palsy. The clusters are given below:

Cluster 1: P4, **P7**, P3, P9 (Number of symptoms present = 1)
Cluster 2: P12, **P18**, P10, P6 (Number of symptoms present = 5)
Cluster 3: P1, P8, P2, **P5**, P13, P16, P20 (Number of symptoms present = 4)
Cluster 4: P15, **P17**, P19 (Number of symptoms present = 4)
Cluster 5: **P14**, P11 (Number of symptoms present = 5)

The bolded components that are selected are representative components or K-Medoids and represent the cluster; the number of symptoms in the concerned component is also specified in the parenthesis.

6. Coupling metric

The components of a dataset would be more error-free if its work is evenly distributed among its own sub-components. Coupling is the extent to which various sub-components interact with one another (Gui and Scott, 2006; Khan et al., 2007). Heavy coupling is never desirable because it makes the sub-components highly interdependent, and any changes in one of them would wreak havoc on the total component. Measurements within couplings are important in the sense that they determine the result of heavy or loose coupling, thus specifying the reliability and efficiency of the classification of the concerned work. At this juncture, it was determined that the Fenton and Melton Software Metric (Alghamdi, 2008) can be used for coupling measurement.

The metric is given by $C(x,y) = \frac{i+n}{n+1}$

where n = number of interconnections between x and y, and i = the level of the highest (worst) coupling type found between x and y.

Table 3 Output of Sokal and Mitchener's simple matching coefficient on the dataset.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
P1	1	0.82	0.64	0.82	1	0.36	0.55	0.82	0.64	0.36	0.64	0.36	0.82	0.73	0.55	1	0.36	0.55	0.64	1
P2	0.818	1	0.64	0.64	0.82	0.55	0.55	1	0.64	0.55	0.45	0.36	1	0.73	0.55	0.82	0.36	0.55	0.64	0.8
P3	0.636	0.64	1	0.64	0.82		0.73	0.64	0.64	0.55	0.45	0.73	0.64	0.36	0.45	0.64	0.45	0.73	0.45	0.6
P4	0.818	0.64	0.64	1	0.64	0.18	0.91	0.64	0.82	0.55	0.45	0.73	0.64	0.55	0.64	0.64	0.64	0.55	0.64	0.6
P5	1	0.82	0.36	0.64	1	0.36	0.55	0.82	0.64	0.36	0.64	0.36	0.82	0.73	0.64	1	0.64	0.55	0.64	1
P6	0.364	0.55	0.36	0.18	0.36	1	0.27	0.36	0.36	0.64	0.55	0.45	0.45	0.45	0.36	0.36	0.36	0.64	0.36	0.4
P7	0.545	0.55	0.73	0.91	0.55	0.27	1	0.55	0.91	0.45	0.36	0.82	0.55	0.45	0.55	0.55	0.55	0.64	0.55	0.5
P8	0.818	1	0.64	0.64	0.82	0.36	0.55	1	0.64	0.55	0.45	0.36	1	0.73	0.64	0.82	0.64	0.55	0.64	0.8
P9	0.636	0.64	0.64	0.82	0.64	0.36	0.91	0.64	1	0.36	0.45	0.73	0.64	0.55	0.64	0.64	0.64	0.55	0.64	0.6
P10	0.364	0.55	0.55	0.55	0.36	0.64	0.45	0.55	0.36	1	0.55	0.64	0.55	0.45	0.55	0.36	0.55	0.45	0.55	0.4
P11	0.636	0.45	0.45	0.45	0.64	0.55	0.36	0.45	0.45	0.55	1	0.36	0.45	0.73	0.64	0.64	0.64	0.55	0.64	0.6
P12	0.364	0.36	0.73	0.73	0.36	0.45	0.82	0.36	0.73	0.64	0.36	1	0.36	0.27	0.36	0.36	0.36	0.64	0.36	0.4
P13	0.818	1	0.64	0.64	0.82	0.45	0.55	1	0.64	0.55	0.45	0.36	1	0.73	0.64	0.82	0.64	0.55	0.64	0.8
P14	0.727	0.73	0.36	0.55	0.73	0.45	0.45	0.73	0.55	0.45	0.73	0.27	0.73	1	0.73	0.73	0.73	0.45	0.73	0.7
P15	0.545	0.55	0.45	0.64	0.64	0.36	0.55	0.64	0.64	0.55	0.64	0.36	0.64	0.73	1	0.64	1	0.18	1	0.6
P16	1	0.82	0.64	0.64	1	0.36	0.55	0.82	0.64	0.36	0.64	0.36	0.82	0.73	0.64	1	0.64	0.55	0.64	1
P17	0.364	0.36	0.45	0.64	0.64	0.36	0.55	0.64	0.64	0.55	0.64	0.36	0.64	0.73	1	0.64	1	0.18	1	0.6
P18	0.545	0.55	0.73	0.55	0.55	0.64	0.64	0.55	0.55	0.45	0.55	0.64	0.55	0.45	0.18	0.55	0.18	1	0.18	0.5
P19	0.636	0.64	0.45	0.64	0.64	0.36	0.55	0.64	0.64	0.55	0.64	0.36	0.64	0.73	1	0.64	1	0.18	1	0.6
P20	1	0.82	0.64	0.64	1	0.36	0.55	0.82	0.64	0.36	0.64	0.36	0.82	0.73	0.64	1	0.64	0.55	0.64	1

Table 4 Single symptom couplings.

Sl. no.	Symptom	Count	Patient IDs
1	3	3	6,10,12
2	8	4	6,11,14,18
3	10	5	6,7,9,12,18
4	9	6	3,6,10,11,12,18
5	11	6	6,10,11,15,17,19
6	2	8	1,5,11,14,15,16,17,20
7	5	9	2,6,8,10,13,14,15,17,19
8	4	10	1,2,3,5,6,8,13,16,18,20
9	7	12	1,2,5,6,8,10,11,13,14,16,18,20
10	6	14	1,2,5,6,8,9,11,13,14,15,16,17,19,20

Table 5 Four symptoms coupling.

Sl. no.	Count	Patient IDs
1	1	6
2	2	10,12
3	3	4,7,9
4	4	1,5,16,20
5	5	2,3,8,13,18
6	5	11,14,15,17,19

Table 6 Total symptoms couplings.

Sl. no.	Count	Patient IDs
1	3	2,8,13
2	3	15,17,19
3	4	1,5,16,20

Considering the *total* dataset that is specified in Table 1, the following results are produced before the metric is applied. Note that this coupling is based on the similarity of the symptoms between the patients.

1. Statistics of single symptom coupling are given in Table 4:
2. Statistics of the first four symptom coupling are given in Table 5:
3. Statistics of all symptom coupling are given in Table 6:

Table 7, given below, specifies only the representative components for each of the clusters. Between these components, the Fenton–Melton Software Metric is applied to measure the coupling.

Table 8 Output of the coupling metric.

	P5	P7	P14	P17	P18
P5	11	1.83	2.75	2.2	1.83
P7	1.83	11	1.57	1.83	2.2
P14	2.75	1.57	11	2.75	1.57
P17	2.2	1.83	2.75	11	1.1
P18	1.83	2.2	1.57	1.1	11

The results are shown below in Table 8:

Evidently, after this carefully planned process, the cluster classification must be the significant output. From Table 8, we can classify the medoids themselves into three clusters: C1 (which consist of P5, P17); C2 (which consist of P14, P18) and C3 (which consists of only P7). Note that the classification has been performed by considering the number of symptoms that are in common. Expanding the medoids into their base clusters, we obtain:

C1: P1, P8, P2, **P5**, P13, P16, P20, P15, **P17**, P19 (4 symptoms)
C2: **P14**, P11, P12, **P18**, P10, P6 (5 symptoms)
C3: P4, **P7**, P3, P9 (1 symptom)

Apparently, it can be concluded that the patients in C2 are suffering with paralysis; the patients in C1 are more likely to become disease-prone in the future, and the patients in C3 are in normal condition. If data on a new patient are offered now, he/she can be easily classified into one of the concerned clusters and can be provided the same medicines given to the previous similar patients.

7. Reliability

The discussed methodology of clustering, coupling and placing a new patient in one of the categories can be applied more confidently if its reliability is also discussed and revealed. In this short exposure, the values of T-wave alternans (TWA) (Wave alternans, 2013) from an ECG are considered for the patients who are shown to be disease-prone (C1). Table 9 consists of the patients, their disease linkages (binary) and their T-wave values.

In this case, Spearman's Rank Correlation Coefficient (2013) is considered and is given as $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$ where d is the distance and n is the number of elements.

In this case, $n = 4$ and the rank correlation can be calculated as 0.8 as per the formula. The resulting values are shown in the table. Because the rank is within admissible limits, it can

Table 7 Representative components for the clusters.

Patient ID (↓)	Alcoholism (S1)	ALS (S2)	Genetic Problems (S3)	Botulism (S4)	Cauda Equina (S5)	Cerebral Palsy (S6)	Coxsackie Virus (S7)	Nerve Problems (S8)	Guillian–Barre Syndrome (S9)	High Blood Potassium (S10)	Stroke (S11)
P5	0	1	0	1	0	1	1	0	0	0	0
P7	0	0	0	0	0	0	0	0	0	1	0
P14	0	1	0	0	1	1	1	1	0	0	0
P17	0	1	0	0	1	1	0	0	0	0	1
P18	0	0	0	1	0	0	1	1	1	1	0

Table 9 Clustered patients with rank correlation factors.

Patient ID	Disease-Prone (A)	T-Wave (B)	A-B (d)	d^2
P6	1	1	0	0
P10	1	0	1	1
P11	1	1	0	0
P12	1	0	1	1
P14	1	1	0	0

be stated that there exists a correlation between the observed clustered categories and the ECG data (T-Wave). It is, hence, concluded here that the proposed classification of patients and medicines offered to the new patient who is nearer to a certain cluster is reliable and can be applied.

8. Results and conclusions

The results obtained from the developed method help to categorize the patients into homogeneous groups and the results obtained from the concerned groups are provided in Table 3. To confirm the disease, each step/medicine that is given to a patient will be of very crucial importance. Hence, the coupling methods of Fenton and Melton (Alghamdi, 2008) have been considered with utmost care for identifying the most relevant patient with the specific symptom. The number of similar patients with similar symptoms is considered to allow them to be sent to a specific specialist. Note here that before ratifying the disease, the reports are also contemplated.

A novel methodology for software reuse of data from the medical domain is presented in this work. The symptoms that pertain to the disease of paralysis are considered in patients from the remote village Chintapalli, where only a primary health center with a technician and a doctor is available, with no specialized aid for treating the local patients. The dataset is generated from these patients, and the general symptoms that pertain to the considered disease of paralysis are considered for this study.

The clustering is performed based on the Pearson Type-VII distribution and the patients are classified into categories. Coupling metrics are used to bundle the patients who have similar symptoms into groups. The Fenton and Melton metric is very convenient for this purpose. The results are duly analyzed and a final clustering is determined to fit all of the patients into the final resulting clusters. The reliability of the medicines that are suggested for a new patient where the same medicines were used previously for patients in the same cluster, is also presented, to make the conclusions more robust.

With the suggestion of medicine that evidently has more of a basis, the confidence of the concerned paramedics or the users will surely be satisfactory when the approach suggested here is used in medicine. Apparently, this work is aimed to be used in remote areas with less availability of doctors, and a developed app can surely be generated for the purpose of saving precious lives of ailing patients at critical junctures.

References

Alghamdi, J., 2008. Measuring software coupling. Arabian J. Sci. Eng. 33 (1B).

Bhanu Sridhar, M., Srinivas, Y., Krishna Prasad, M.H.M., 2012. Software reuse in cardiology related medical database using K-means clustering technique. J. Software Eng. Appl. 5.9, 682–686.

BhanuSridhar, M., Srinivas, Y., Krishna Prasad, M.H.M., 2012. Software Reuse in Medical Database for Cardiac Patients using Pearson Family Equations 58 (14).

Maria C., Nascimento et al., A Hybrid Heuristic for the K-Medoids Clustering Problem, GECCO '12, July 7–11, 2012, Philadelphia, Pennsylvania, USA.

Christopher & Dana Reeve Foundation, Paralysis Resource Center, <http://www.christopherreeve.org/site/c.mtKZKgMWKwG/b.5184189/k.5587/Paralysis_Facts__Figures.html>, 21/8/2012.

Disease datasets from Machine Learning Repository, <http://archive.ics.uci.edu/ml/datasets>, 21/8/2012.

Godin, Robert et al, 1995. Applying concept formation methods to software reuse. IJSEKE 5 (1), 119–142.

G. Gui, P.D. Scott, Coupling and Cohesion Measures for Evaluation on Component Reusability, MSR '06, May 2006, Shanghai, China. <http://en.wikipedia.org/wiki/Categorical_data>, 20/4/2013.

http://en.wikipedia.org/wiki/Expert_system, Wikipedia, 20/4/2013.

Khan R.A., et al., An Empirical Validation of Object Oriented Design Quality Metrics, Journal of King Saud University – Computer and Information Sciences, 2007, Vol. 17, Riyadh.

Liao, Shu-Hsien, 2004. Expert System Methodologies and Applications – A Decade Review from 1995 to 2004. ELSEVIER.

McLachlan, Geoffrey, Peel, David, 2000. In: Finite Mixture Models, 299. Wiley-Interscience.

Medline Net, <http://www.medicinenet.com/paralysis/symptoms.html>.

Medline Plus, <http://www.nlm.nih.gov/medlineplus/paralysis.html>, 22/8/2012.

Patil, Shantakumar, Kumaraswamy, Y.S., 2009. Intelligent and effective heart attack prediction system using data mining and artificial neural network. Eur. J. Sci. Res. 31 (4).

Paul, Razan, Abu Sayed Md Latiful Hoque. "A storage & search efficient representation of medical data". Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on. IEEE, 2010.

David, Peel, McLachlan, Geoffrey J., 2000. Robust mixture modelling using the *t* distribution. Statistics and Computing 10 (4), 339–348.

Spearman's Rank Correlation Coefficient, Rank Correlation, <http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>, 20/4/2013.

Řezanková, H., 2009. Cluster analysis and categorical data. Statistika, 216–232.

Sun, Jianyong, Kabán, Ata, Garibaldi, Jonathan M., 2010. Robust mixture clustering using Pearson type VII distribution. Pattern Recogn. Lett. 31 (16), 2447–2454.

T Wave alternans, <http://en.wikipedia.org/wiki/T_wave_alternans,20/4/2013>.