REVIEW ARTICLE

Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies

Hamid Reza Marateb[1], Marjan Mansourian[2], Peyman Adibi[3], Dario Farina[4]

[1]Department of Biomedical Engineering, Engineering Faculty, the University of Isfahan, Isfahan, Iran, [2]Department of Biostatistics and Epidemiology, Health School, Isfahan University of Medical Sciences, Isfahan, Iran, [3]School of Nutrition and Food Science, Isfahan University of Medical Sciences, Isfahan, Iran, [4]UNIVERSITÄTSMEDIZIN GÖTTINGEN, GEORG-AUGUST-UNIVERSITÄT, Department of Neurorehabilitation Engineering, Bernstein Focus Neurotechnology Göttingen, Bernstein Center for Computational Neuroscience, Göttingen, Germany

Background: selecting the correct statistical test and data mining method depends highly on the measurement scale of data, type of variables, and purpose of the analysis. Different measurement scales are studied in details and statistical comparison, modeling, and data mining methods are studied based upon using several medical examples. We have presented two ordinal–variables clustering examples, as more challenging variable in analysis, using Wisconsin Breast Cancer Data (WBCD). **Ordinal-to-Interval scale conversion example:** a breast cancer database of nine 10-level ordinal variables for 683 patients was analyzed by two ordinal-scale clustering methods. The performance of the clustering methods was assessed by comparison with the gold standard groups of malignant and benign cases that had been identified by clinical tests. **Results:** the sensitivity and accuracy of the two clustering methods were 98% and 96%, respectively. Their specificity was comparable. **Conclusion:** by using appropriate clustering algorithm based on the measurement scale of the variables in the study, high performance is granted. Moreover, descriptive and inferential statistics in addition to modeling approach must be selected based on the scale of the variables.

Key words: Biostatistics, breast cancer, cluster analysis, data mining, research design

How to cite this article: Marateb HR, Mansourian M, Adibi P, Farina D. Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. J Res Med Sci 2014;19:47-56.

INTRODUCTION

In medical research, the design of a study is the most important part that directs other steps of research, especially, all type of data analysis. A badly designed study could never be retrieved, whereas a poorly analyzed one can usually be re-analyzed.[1] Another important issue, such as sample size calculation, also depends on the kind of experimental design and kind of measurements that exist in the study. Above all, the main question is: What types of data are being measured? The other steps of the analysis are indeed determined by the type of variable used.[2-6] In this regard, analyzers assume that the variables have specific levels of measurement.

Stevens proposed his typology in 1946.[7] In his article, Stevens claimed that all measurements in science were conducted using four types of scales that he called 'nominal', 'ordinal', 'interval' and 'ratio', unifying both qualitative (which are described by his 'nominal' type) and quantitative (to a different degree, all the rest of his scales). The concept of scale types later received the mathematical rigor that it lacked at its inception with the work of mathematical psychologists Theodore Alper,[8, 9] Louis Narens,[10, 11] and R. Duncan Luce.[12-14] Nowadays, the ordinal scale is considered as a qualitative variable.[15] However, this scale typology has received a lot of criticism.[6, 16-18] Alternative scale taxonomies have therefore been suggested[19] that consists of grades, ranks, counted fractions, counts, amounts, and balances.[6] Most of the conflict between the pro-Stevens ('conservative') and the anti-Stevens ('liberal') camps begins after both sides agree that a certain variable is ordinal. But they part company when analyzing the data generated by that variable. The exchange in Nursing Research between Armstrong and Knapp is illustrative of the competing positions.[20]

Measurement scales

Nominal scales are only used for qualitative classification. They can be only measured whether the individual items belong to certain distinct categories. However, it is not possible to quantify or rank order the categories. Nominal data has no order, and the categories assignment is arbitrary. Also, it is not possible to

Address for correspondence: Dr. Marjan Mansourian, Department of Biostatistics and Epidemiology, Health School, Isfahan University of Medical Science, Isfahan, Iran. E-mail: j_mansourian@hlth.mui.ac.ir

Received: 21-10-2013; **Revised:** 11-11-2013; **Accepted:** 24-11-2013

perform arithmetic or logical operations on the nominal data.[18] Briefly, nominal data have three distinct features: 1) no ordering of the different categories, 2) no measure of distance between values, and 3) categories can be listed in any order without affecting the relationship between them. Nominal variables are also called (nonranked) categorical in the literature. The number of occurrences in each category is referred to as the frequency count for that category.[6] The other category dichotomous (binary) is defined as the variables that are nominal variables that have only two categories or levels. Examples of normal variable are gender, marital status, eye color, nationality, affiliation, religious preference, surgical outcome (dead/alive), blood type, and epidemiological status (healthy, patient), having any symptoms in a questionnaire (yes/no).

A discrete–ordinal scale is a nominal variable, but the different states are ordered in a meaningful sequence. Ordinal data have order, but the intervals between scale points may be uneven. Because of the lack of equal distances, arithmetic operations are not possible, but logical operations can be performed.[21] Under an ordinal scale, the subjects or objects are ranked in terms of degree to which they possess a characteristic of interest.[6] An ordinal scale indicates direction, in addition to providing nominal information. In medicine, ordinal variables often describe the patient's characteristics, attitude, behavior, or status. Examples of ordinal variables might include: stages of cancer (stage I, II, III, IV), education level (elementary, secondary, college), pain level (1-10 scale), satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied), social status (upper, middle, lower), type of degree (BS, MS, PhD), the Likert variable[22] such as the attitudinal response variable (agreement level) with four levels (strongly disapprove, disapprove, approve, strongly approve), or 4-item–rating scale (always, often, sometimes, never), graduation rank, visual analog scale (VAS), BMI (body mass index)-based nutritional status (sever thin, thin, normal, overweight, and obese).

Continuous — ordinal scales occur when the measurements are continuous, but one is not certain whether they are on a linear scale, the only trustworthy information being the rank order of the observations. For example, if a scale is transformed by an exponential, logarithmic, or any other nonlinear monotonic transformation, it loses its interval scale property. Here, it would be expedient to replace the observations by their ranks.[21]

Interval scales are metric scales that have constant, equal distances between values, but the zero point is arbitrary. They are measured on a linear scale, and can take on positive or negative values. It is assumed that the intervals keep the same importance throughout the scale.[21] In an interval scale, such as body temperature (ºC, ºF) or calendar dates, a difference between two measurements has meaning, but their ratio does not.[23] Counts are interval scale measurements, such as counts of publications or citations, years of education, intelligence (IQ test score), BMI, and age (years).

The ratio scales are metric scales and the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. Briefly, ratio scales have equal intervals between values, the zero point is meaningful, and the numerical relationships (e.g. division) between numbers are meaningful. Examples of the ratio scales include weight, pulse rate, respiratory rate, body temperature (ºK), and body length in infants or height in adults. Since the statistical tests on the ratio scales are the same as those of interval scales, the inferential statistics will be discussed on normal, ordinal, and interval scales.

Statistics are part of our everyday life. Anyone who lacks fundamental statistical literacy, reasoning, and thinking skills might not be able to perform acceptable research. Kuzma provided a formal definition of the term 'statistics':[24]

'A body of techniques and procedures dealing with the collection, organization, analysis, interpretation, and presentation of information that can be stated numerically'. The statistical analysis divided in two important branches; descriptive and inferential analysis.

Descriptive and inferential statistics for different types of variables

Descriptive statistics is the strategy of quantitatively describing the main features of a collection of data and presented by central and dispersion tendencies. The central tendency of nominal variables is defined as the mode, the most common item. For the ordinal variables, the median (middle-ranked item), or the mode can be used as the central tendency estimates. For interval variables, the mode, median, and arithmetic mean could be used as the central tendency, yet in addition to the aforementioned operators, the geometric (the samples root of the product of the data samples) and harmonic (the reciprocal of the arithmetic mean of the reciprocals of the data samples) means are allowed for ratio variables.

Statistical dispersion is not defined for nominal and ordinal scales. For interval variables, the range, and standard deviation could be used as the dispersion measure, yet in addition to the aforementioned operators, the studentized range (the difference between the largest and smallest data, divided by the standard deviation) and the coefficient of variation (the ratio of the standard deviation to the mean)

are allowed for ratio variables. The inferential statistics used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems is affected by random variation. Any statistical inference requires some assumptions. Rejection of a hypothesis is an important part of inferential statistics using suitable statistical tests as parametric or nonparametric. In parametric tests, the probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters whereas in nonparametric tests the assumptions made about the process generating the data are much less than in parametric statistics and may be completely undefined. The purpose of the analysis and the scale of the measurement of the data define the suitable statistical test.[4] Usually, the statistical parametric tests rely on the normality of the distribution of the interval-scale data. Thus, normality tests such as Kolmogorov–Smirnov or Shapiro–Wilks are used to check the normality assumption.[25] The power of the parametric tests is higher than the corresponding nonparametric tests. Thus, the transformation of the interval variables is sometimes used to guarantee normality assumption.[26]

The appropriate tests for different variable scales for comparisons between two or more groups containing independent or paired samples are listed in [Table 1]. The following clinical examples are given to elaborate the issue of correct statistical test to use the following.

- Comparing the HDL (High-density lipoprotein) value in the healthy and diabetic patients, two independent sample t-test is used if HDL values are normally distributed in the classes, otherwise Wilcoxon–Mann Whitney test is used.
- To identify whether gender is equally distributed among abdominal obese people, the Chi-square test can be used.
- If the distribution of the BMI-based nutritional status (sever thin, thin, normal, overweight and obese) is the same among the patients with liver cancer, the Wilcoxon–Mann Whitney test is used.
- Finding whether the prevalence of high diastolic pressure is the similar in the Normoalbuminuria, Microalbuminuria, and Macroalbumineria groups, the Chi-square test could be used.
- The effectiveness of an educational program on the correct diagnosis of a disorder is identified using the McNemar test.
- The difference of blood sample vitamin-D concentration in normal, pre-diabetic and diabetic patients is identified using one-way ANOVA.
- The comparison of blood HbA1C concentration among pregnant women in the first, second and third semester of the pregnancy is performed using one-way repeated measurements ANOVA.

Table 1: Selecting the appropriate test for comparisons between two or more than two groups based on different scales

Scale of variables	Number of sample comparison	Independent samples test(s)	Paired samples test(s)
interval & normal	2	2 independent sample t-test	paired *t*-test
ordinal or interval	2	Wilcoxon-Mann Whitney test	Wilcoxon signed ranks test
Categorical (binary)	2	Chi-square test or Fisher's exact test	McNemar test
interval & normal	>2	one-way ANOVA	one-way repeated measurements ANOVA
ordinal or interval	>2	Kruskal Wallis test	Friedman test
categorical	>2	Chi-square test	Generalized estimating equations

Additionally, appropriate modeling methods for different variable scales are listed in [Table 2]. Modeling is usually used when we want to reduce the effect of confounders and the type of the modeling is determined by the scale of the dependent variable(s). Here are some clinical modeling examples.

- The gender-specific difference of blood sample vitamin-D concentration in normal, pre-diabetic and diabetic patients is identified using factorial ANOVA.
- The effect of air pollutant concentration on the born weight considering mother's nutritional status and the supplementary intake is determined using the multiple linear regression.
- In the later example, if the born weight is categorized by the underweight and normal groups, the simple logistic regression is used.
- The effectiveness of a treatment method on stage of tumor (grades I–IV), cancelling the effect of confounders such as gender, age, and immunologic factors of patients is determined by using ordered logistic regression.

For detailed description of the aforementioned methodologies the reader is referred to the selected textbooks and guidelines.[4, 6, 27-30]

Data mining for different types of variables

Data mining (DM) is the process of discovering new patterns embedded in large data sets. DM uses this information to build predictive models. A lot of complex data are generated by healthcare systems in which manual analysis has become impractical. DM can generate information that can be useful to health care, including patients by identifying effective treatments. DM of medical data requires specific medical and DM knowledge. Medical DM activities include clustering, classification and estimation, and treatment effectiveness.[31-33] In this section, we focus on clustering.

Table 2: Selecting the appropriate test or modeling for different categories of dependent and independent variables

Number of dependent variable	Number of independent variable	Scale of independent variable	Scale of dependent variable	Test or kind of modeling
1	≥2	Categorical	interval & normal	factorial ANOVA
1	≥2	Every scale	ordinal or interval	ordered logistic regression
1	≥2	Every scale	Categorical	Multinomial logistic regression
1	1	Interval	interval and normal	correlation or simple linear regression
1	1	Interval	ordinal or interval	non-parametric correlation
1	1	Interval	Categorical	simple logistic regression
1	≥1	Interval or categorical	interval and normal	multiple linear regression or analysis of covariance (ANCOVA)
1	≥1	Interval or categorical	Categorical	multiple logistic regression or discriminant analysis
≥2	1	Categorical	interval and normal	one-way MANOVA
≥2	≥2	Every scale	interval and normal	multivariate multiple linear regression

However, the issues considered can be extended to other DM methods.

Clustering is the task of grouping a set of objects in such a way that objects belonging to the same cluster are similar to each other (homogeneity) and objects belonging to different clusters are dissimilar to each other (separation). A clinical example is now given for clarification of clustering procedure: in year 2000, a paper was published in Nature by Alizadeh *et al*.,[34] in which the gene expression profiles (micro array) of 72 patients diagnosed as either acute myeloid leukemia (AML) or acute lymphatic leukemia (ALL) were analyzed. The authors could distinguish two similar groups corresponding to AML and ALL by clustering and match the groups with the routine leukemia diagnosis. Based upon this Roland Eils designed an expert system for prediction of genetic disease.[35] In the other words, if a new microarray gene profile is tested, it is possible to diagnose type of leukemia.

The similarity between objects plays an important role in any clustering algorithm, since similar objects belong to a cluster. An object could be a patient with variety of recorded clinical data (features). Similar objects have similar features. Features could be interval, ordinal, and nominal variables. The question is how the similarity is measured for various types of data scales?

The dissimilarity measure (distance) can be easily defined for interval variables. The Euclidean, Manhattan, Maximum, Minkowski, Mahalonobis, Average, Chord, Canberra, and Czekanowski distances could be used in this case.[36] For the nominal variables, simple matching, Russell-Rao, Jaccard, Dice, Rogers-Tanimoto, and Kulczynski distances might be used, while there are more than 76 distance measures such as Yule, Sokal-Sneath-c, and Hamann measures that could be used for the binary data.[36-38] An example is shown in [Figure 1] for better clarification. However, there are many problems in defining dissimilarity measures for ordinal

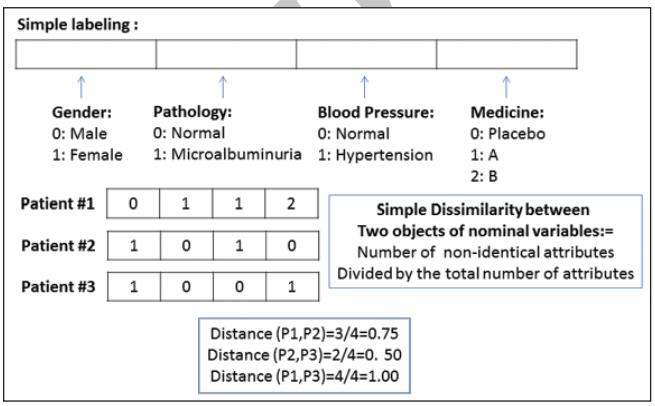

Figure 1: An example of calculating the distance between two objects of ordinal variables, using the simple dissimilarity measure

variables. The distance measure for the ordinal data cannot be defined unless the ordinal to interval variable conversion is used. Moreover, defining proper similarity measure can also affect statistical feature reduction and visualization techniques such as multidimensional scaling (MDS), in which the distance measure is defined for different measurement scales (e.g. using the weighted Euclidean model).[39-42]

Ordinal to interval variable conversion

Consider the four-item rating scale (always, often, sometimes, never) that is widely seen in the questioners of psychological,[43] gastrointestinal,[44] nutritional,[45] and public health[46] researches. One approach to handle ordinal variables is introducing a dummy binary variable by merging [always, sometimes] and [rarely, never] as 'yes' or 'no'. Thus, the ordering information is discarded and a suitable binary distance measure can be used. However, some information is lost, that could have potentially improved the predictive performance of the groups' dissimilarity.[47]

The other strategy is monotonic nonrandom and random assignments of numbers to rank order and treat them as

if they conform to interval scale.[48, 49] The first approach is called equal distance scoring (EDS), while the other solution is entitled as monotonic random scoring (MRS) in the literature. Using EDS, interval variables such as [0, 1, 2, and 3] are used for the four-item rating scale. Accordingly, the distance between 'sometimes' and 'never' is the same as that of 'sometimes' and 'often'. This is not really correct. Additionally, EDS has received criticisms in the literature and proved not to be efficient even in correlation analysis in some cases where the ranks are not uniformly distributed.[50] Although, MRS has been extensively used in the literature, it has also received criticisms.[51] In MRS, uniform and normal monotonic random numbers are generated and used instead of the ordinal scale. Using MRS, the aforementioned four-item rating scale might be represented by the following uniform monotonic random numbers [0.1270, 0.8147, 0.9058, and 0.9134]. Using the random number generator again, the new mapping would be [0.0975, 0.2785, 0.5469, and 0.6324]. The question is whether the transformation is unique at every MRS run, and also if the problem mentioned in EDS is resolved?

The optimal ordinal-to-interval conversion is still debatable and many complicated approaches have been introduced in the literature.[51, 52] In none of which, the mapping was not defined as to maximize the separation of the groups in the clustering procedure. In the next section, clustering methods defined for different variable scales are discussed and the relationship between this mapping and clustering is considered.

Clustering methods for different variable scales

Most previous clustering methods focus on interval data for which the dissimilarity could be calculated easily, such as density-based (DBSCAN,[53] OPTICS[54]), partitioning (k-means,[55] k-medoids,[56] fuzzy c-means,[57] ISODATA[58]), hierarchical (different linkage algorithms,[59, 60] MONA,[61] DIANA[62]), and grid-based (WaveCluster,[63] Fractal Clustering[64]).

Nonranked categorical clustering algorithms have been extensively proposed in the literature, such as LIMBO,[65] COOLCAT,[66] CACTUS,[67] ROCK,[68] MMR,[69] CLICKS,[70] HD vector,[71] AUTOCLASS,[72] K-modes,[73] fuzzy K-modes,[74] fuzzy centroids,[75] genetic fuzzy k-modes,[76] and fuzzy centroids.[75] However, the dissimilarity measures and cluster representatives have great impact on the clustering performance and convergence.[77-79]

It is possible to use dummy binary variables for ordinal data, and then use any of the above clustering methods at the expense of losing details. There are few algorithms proposed for clustering ordinal data, such as median fuzzy c-means[80] and a modified fuzzy c-means clustering method in which the ordinal-to-interval mapping is simultaneously determined by particle swarm optimization.[81] In the later method, the mapping is calculated so as to maximize the inter-cluster distance and minimize the intracluster distance. This algorithm is one of the few clustering methods in which the mentioned transformation is adaptively estimated for each ordinal variable. This algorithm will be used at the next section of this manuscript for clustering a cancer dataset with ordinal variables.

Latent variable models

Latent variable models, specifically item response theory, have also been used for modeling and clustering of ordinal data.[82-84] The mixture of item response models could be used for the clustering of such data. It is assumed that the observed ordinal data are discrete versions of an underlying latent Gaussian variable. The clustering is then achieved by fitting a mixture model to the latent Gaussian data.[85] However, this method relies on the posterior mean of the latent Gaussian data and the Gaussian assumption could be valid for a sufficiently large data set (number of variables and also levels of ordinal variable) which cannot be always taken for granted.[85]

Latent class analysis

Latent class analysis (LCA) is a subset of structural equation modeling, used to find groups or subtypes of cases in multivariate categorical data. These subtypes are called 'latent classes'.[86] One of the common statistical application areas of LC analysis is the clustering, in which LC cluster models are introduced. These models have advantages over traditional clustering methods: such as probability-based classification (similar to fuzzy memberships), handling continuous, categorical, counts,[87] or mixed mode data[88-90] and the application of demographics and other covariates for clustering analysis.[91-94] LC models are model-based clustering methods in which explicit assumptions are made about the form of the probability density function describing the population of the observed data.[95, 96] Clustering analysis and further inferences about the numbers of clusters and cluster membership are based on estimation of the unknown parameters in the probability model used.[97] Two main methods to estimate the parameters of the various types of LC cluster models are the maximum-likelihood (ML) method and the maximum-posterior (MAP) method; thus, a well-known problem in LC analysis is the occurrence of local solutions. Accordingly, the analyst must interpret estimates cautiously. Moreover, the weak identifiability of LC clustering,[98] the complexities of the likelihood function and likelihood surface make the procedure sensitive to initial estimates.[99] Also, the model selection issue is one of the main research topics in LC clustering, that is, estimation of the number of clusters and the form of the model given the number of clusters. Akaike (AIC), Bayesian (BIC), and

consistent Akaike (CAIC) information criteria have been used for model selection.[100] Software packages such as MCLUST,[101] Mplus,[102] poLCA,[103] Latent GOLD,[104] and SAS[99] can be used for LC cluster analysis.[105]

Mixed data

In many applications, each instance in a data set is described by more than one type of attribute. For example, we would like to group people based on their recorded anthropometric or clinical data. This grouping can identify different diseases. The recorded data for each person contain gender (binary variable), the assignment to (underweight, normal, overweight, and obese classes) (ordinal variable), HDL and LDL cholesterol values (interval), etc. This is an example of mixed-type data, in which similarity and dissimilarity between two instances (e.g. people) cannot be calculated using the methods discussed so far. A general distance coefficient and a generalized Minkowski distance was introduced for mixed-type data in the literature.[36] Other methods have also been introduced in the literature.[106-112]

ORDINAL-TO-INTERVAL SCALE CONVERSION EXAMPLE

Since there are few studies on ordinal data clustering, an example is given based on the breast cancer databases obtained from the Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)). This database, known as Wisconsin Breast Cancer Data (WBCD) with the number of web hit of 98032, was obtained from the university of Wisconsin Hospitals, Madison by Dr. William H. Wolberg[113-116] and has been extensively used as a clustering benchmark in the literature.[81, 117, 118] There are 699 patient records in the database. Each attribute has 10 ordinal values. Sixteen patient recordings had missing values, excluded. Thus, the sample size was 683. Each recording represents nine measurements made on a fine needle aspirate (FNA) taken from the patient breast. The nine cytological measurements are the clump thickness, size uniformity, shape uniformity, marginal adhesion, cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. Each of these measurements are described by an ordinal integer label between 1 and 10, the larger the number the greater likelihood of malignancy.[115] These ratings were done by the clinical experts. All malignant aspirates were histologically confirmed whereas FNAs diagnosed as benign masses were biopsied only at the patient's request. The remainder of benign cytologies was confirmed by clinical re-examination 3 and 12 months after the aspiration. Masses that produced unsatisfactory or suspicious FNAs were surgically biopsied.[114] Accordingly, 239 cases were diagnosed as malignant and 444, as benign. The class labels were saved as the gold standard and kept for comparison. The class labels were excluded from the data set; thus 683 10-dimensional ordinal dataset was used for clustering. The number of clusters (groups) was estimated and the accuracy of malignant and benign classification was assessed by comparison with the gold standard.

Since ordinal data clustering is more challenging than clustering other types of data, we consider two different ordinal clustering methods for analyzing WBCD. The first approach was taken from the literature while the second one is proposed by the authors of this manuscript.

Ordinal data clustering based on modified FCM analysis (clustering #1)

Using the ordinal dataset, a modified fuzzy c-means whose ordinal-to-interval conversion was estimated based on the particle swarm optimization was used.[81] The algorithm was run from 2 to 10 numbers of clusters, and the clustering structure with optimum Xe-Beni clustering validity index[119] was selected. In the other words, number of clusters with better relative compactness (minimum intra-cluster distance) and separation (maximum intercluster distance) was chosen.[120] In the selected clustering structure, the malignant and benign clusters were identified by comparison with the gold standard and the errors were reported. Errors included number of malignant cases in the benign cluster and vice versa.

Ordinal data clustering based on modified OPTICS analysis (clustering #2)

The ordinal data were converted to interval data by using the EDS algorithm. It was because the ordinal scales were equally assigned without prior expert-based knowledge. Then, a density-based clustering method OPTICS was used to identify the clustering structure. OPTICS resolves the problem of detecting meaningful clusters in data of varying density such that points that are spatially closest in the multidimensional space become neighbors in the ordering. OPTICS can identify clustering structure, and unlike FCM does not need major input parameters or postprocessing such as clustering validity analysis.[121] Like the previously mentioned clustering method, the malignant and benign clusters were identified by comparison with the gold standard and the errors were reported.

Clustering performance analysis

The values of true positive (TP), true negative (TN), false positive (FP), and false Negative (FN) were calculated for each of the aforementioned clustering methods, by comparing the clustering results with those of the gold standard. Then, the information theory parameters were calculated as the following:

Sensitivity (Se) = Recall (Re) = TP/(TP+FN);
Specificity (Sp) = TN/(FP+TN);
Precision (Pr) = TP/(TP+FP);

Type I error: FP rate (α) =1-Sp;
Type II error: FN rate (β) =1-Se;
Power =1-β =Se;
F-score =2*(Pr*Re)/(Pr+Rl) = harmonic mean (Pr,Rl);
Accuracy (Acc) = (TP+TN)/(TP+TN+FN+FP);

The codes of the above-given two clustering algorithms and the validation program were written in MATLAB (MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States), and is available upon the request to the authors.

RESULTS

In the first clustering method, Xe-Beni index showed the optimum value at two clusters. It showed that there were two clusters in the data, which is quite reasonable. The FCM clustering algorithm was run 10 times, and the clustering results with the best compactness and separation were used.[122] The ordinal-to-interval conversion matrix for nine ordinal variables with 10 ranks was listed in [Table 3]. The ranks of different ordinal variables were transformed differently. In the other words, the transformation was done, so as to optimize the clustering structure. Comparing with the gold standard, the performance of the first clustering method is listed in [Table 4].

Using the clustering method #2 with 40-nearest neighbors (40-NN), the reachability distance plot (RD-plot) was shown in [Figure 2]. This 1D plot shows the clustering structure of the multidimensional data, in which major local minimums correspond with a cluster. In this plot, two major clusters were detected related to malignant and benign groups, respectively. Although the major local minimums could be detected manually, there are methods for automatically detecting including clusters.[121] The performance of this clustering method was shown in [Table 4].

The power of both of clusters methods are 98%, while the type-I error (α) was 0.03 and 0.09 for the clustering methods #1 and #2, respectively. In both of the clustering methods, the FN-rate (β) was 0.02. A FN is much more serious than a FP since it means that the subject will not be treated.[81] Both of aforementioned methods, showed 'almost perfect agreement' with the gold standard.

DISCUSSION

One of the important elements of a good medical research is identifying the key variables of the study and their method of measurement (measurement scale) and unit of measurement.[123] In addition to different types of variables,[124] such as independent (risk factors), dependent (outcome), confounding (intervening), and background variables, the scale of variables (qualitative versus metric) plays an important role of selecting appropriate statistical tests. Due to the importance of selecting appropriate statistical comparison and modeling tests, they have been mentioned in [Tables 1 and 2], in detail. Also, clinical examples taken from different medical studies were given in this paper for better elaboration. Although the selection of appropriate tests have been studied in the manuscripts,[4, 6] this manuscript is one of the first one of its kind to discuss about different variable scales and their suitable statistical and data mining methods with several examples. Much of what was written in the literature is

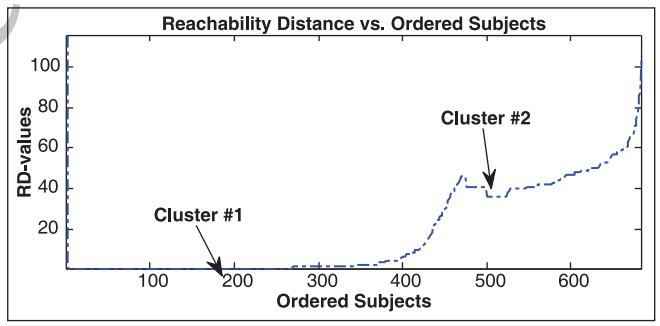

Figure 2: The clustering structures of WBCD, found by the second ordinal–variable clustering method. Each major valley (local minimum) of the reachability distance plot (RD-plot) corresponds with a possible cluster. In this example, the first cluster is the malignant group while the second one is the benign group.

Table 3: The ordinal-to-interval conversion matrix for nine ordinal variables (columns) with 10 ranks (rows) studied on the WBCD using the clustering method #1

Ordinal Rank	1	2	3	4	5	6	7	8	9
1	0.0849	0.0264	0.1216	0.0994	0.1291	0.1451	0.1164	0.1469	0.1177
2	0.1705	0.0737	0.1337	0.1211	0.1627	0.1838	0.214	0.1588	0.1324
3	0.2147	0.1439	0.1366	0.1288	0.2112	0.2352	0.2493	0.2160	0.2284
4	0.2974	0.214	0.1773	0.2203	0.2135	0.3097	0.2802	0.2335	0.2691
5	0.323	0.2296	0.2100	0.2404	0.2441	0.3277	0.2946	0.2371	0.3674
6	0.3471	0.3091	0.2930	0.2820	0.3277	0.3663	0.3362	0.2515	0.3916
7	0.3494	0.3196	0.3432	0.3441	0.3578	0.3940	0.4227	0.2783	0.4721
8	0.3915	0.3677	0.3458	0.3468	0.4093	0.4125	0.5091	0.3161	0.5446
9	0.458	0.4185	0.3613	0.3967	0.4328	0.4459	0.516	0.3206	0.5501
10	0.4847	0.4189	0.3621	0.4777	0.5063	0.5086	0.5643	0.3216	0.6769

Table 4: The performance of the clustering methods studied on the WBCD

Clustering method	TP	TN	FN	FP	Se	Sp	Pr	F-score	Acc
#1	437	222	7	17	98	97	96	97	96
#2	435	218	9	21	98	91	95	97	96

WBCD: the Wisconsin Breast Cancer Data; Clustering methods: #1 (Ordinal data clustering based on modified FCM), #2 (Ordinal data clustering based on modified OPTICS); the performance indices: TP: True Positive, TN: True Negative, FN: False Negative, FP: False Positive, Se: Sensitivity, Sp: Specificity, Pr: Precision, F-Score: the harmonic mean of Precision and (Recall = Se), Acc: Accuracy. The detailed description of the above performance indices is mentioned in the clustering performance analysis section.

about clustering analysis and validity analysis of interval data,[62, 120, 125] but little was mentioned about the analysis of categorical variables. In this paper, we discussed about different clustering methods for categorical data and as the first manuscript in review, two different clustering methods were used for analyzing the ordinal WBCD. The first approach was already proposed and tested,[81] while the second approach was proposed by the authors. We hope that this review will be of use for researchers in the field of biomedical sciences.

One of the main limitations of this manuscript is that most of the nominal-data clustering methods were only mentioned and cited. There was no criterion to select in this paper. We have been contacting the authors of the corresponding papers. Most of the clustering programs were received. Some of which were re-compiled in different operating systems, for example, Linux, with the help of other data-mining researchers from different countries. We will be trying to run several clustering algorithms on categorical data on standard Benchmark datasets to have a fair comparison. It will be the focus of our future work.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Sobhan Goudarzi for the implementation of the first ordinal clustering algorithm. This study was supported by the University of Isfahan and Isfahan University of Medical Sciences.

REFERENCES

1. Campbell MJ, D Machin, J Wiley: *Medical Statistics: A Commonsense Approach*. Vol. 2Wiley London, 1993.
2. Swinscow TDV, MJ Campbell: *Statistics at Square One*Bmj London, 2002.
3. Marusteri M, V Bacarea: Comparing Groups for Statistical Differences: How to Choose the Right Statistical Test? *Biochemia medica* 2010; 20: 15-32.
4. McCrum-Gardner E: Which Is the Correct Statistical Test to Use? *British Journal of Oral and Maxillofacial Surgery* 2008; 46: 38-41.
5. McDonald JH: *Handbook of Biological Statistics*. Vol. 2Sparky House Publishing Baltimore, 2009.
6. Lawal B: *Categorical Data Analysis with Sas and Spss Applications* Mahwah, N.J.: Lawrence Erlbaum Associates, 2003; pp. vii, 561 p.
7. Stevens SS: *On the Theory of Scales of Measurement*. Bobbs-Merrill, College Division 1946.
8. Alper TM: A Note on Real Measurement Structures of Scale Type (*M*, *M*+ 1). *Journal of Mathematical Psychology* 1985; 29: 73-81.
9. Alper TM: A Classification of All Order-Preserving Homeomorphism Groups of the Reals That Satisfy Finite Uniqueness. *Journal of Mathematical Psychology* 1987; 31: 135-54.
10. Narens L: A General Theory of Ratio Scalability with Remarks About the Measurement-Theoretic Concept of Meaningfulness. *Theory and Decision* 1981; 13: 1-70.
11. Narens L: On the Scales of Measurement. *Journal of Mathematical Psychology* 1981; 24: 249-75.
12. Luce RD: Uniqueness and Homogeneity of Ordered Relational Structures. *Journal of Mathematical Psychology* 1986; 30: 391-415.
13. Luce RD: Measurement Structures with Archimedean Ordered Translation Groups. *Order* 1987; 4: 165-89.
14. Luce RD: Conditions Equivalent to Unit Representations of Ordered Relational Structures. *Journal of Mathematical Psychology* 2001; 45: 81-98.
15. Mendenhall W, T Sincich: A Second Course in Statistics. 1996.
16. Lord FM: On the Statistical Treatment of Football Numbers. 1953.
17. Guttman L: A General Nonmetric Technique for Finding the Smallest Coordinate Space for a Configuration of Points. *Psychometrika* 1968; 33: 469-506.
18. Velleman PF, L Wilkinson: Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician* 1993; 47: 65-72.
19. Mosteller F, JW Tukey: Data Analysis and Regression. A Second Course in Statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley, 1977* 1977; 1.
20. Knapp TR: Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy. *Nursing Research* 1990; 39: 121-3.
21. Nagpaul P: Guide to Advanced Data Analysis Using Idams Software. *Retrived online from/www. unesco. org/webworld/idams/advguide/TOC. htm S* 2001.
22. Likert R: A Technique for the Measurement of Attitudes. *Archives of psychology* 1932.
23. Campbell MJ, D Machin, SJ Walters: *Medical Statistics: A Textbook for the Health Sciences*Wiley. com, 2010.
24. Kuzma JW: *Basic Statistics for the Health Sciences*. 1st edn Palo Alto, Calif.: Mayfield Pub. Co., 1984; pp. xiv, 274 p.
25. Bland M: *An Introduction to Medical Statistics*Oxford University Press, 2000.
26. Siegel S: Nonparametric Statistics for the Behavioral Sciences. 1956.
27. Pallant J: *Spss Survival Manual: Version 12*Open University Press, 2004.
28. Brown BW: *Beyond Anova: Basics of Applied Statistics*. Vol. 40CRC Press, 1997.
29. Forthofer RN, ES Lee, M Hernandez: *Biostatistics: A Guide to Design, Analysis and Discovery*Academic Press, 2006.
30. Rosner BA: *Fundamentals of Biostatistics*CengageBrain. com, 2011.
31. Lavrač N: Selected Techniques for Data Mining in Medicine. *Artificial intelligence in medicine* 1999; 16: 3-23.
32. Lavrač N, B Zupan: *Data Mining in Medicine*Springer, 2005.
33. Lavrač N, B Zupan: *Data Mining in Medicine*, in *Data Mining and Knowledge Discovery Handbook*, ed. by Maimon O , Rokach LSpringer US, 2005); pp. 1107-37.
34. Alizadeh AA, MB Eisen, RE Davis, C Ma, IS Lossos, A Rosenwald, et al.: Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. *Nature* 2000; 403: 503-11.
35. EILS R: *Expert System for Classification and Prediction of Genetic Diseases*. WO Patent 2,002,047,007 2002.
36. *6. Similarity and Dissimilarity Measures*, in *Data Clustering: Theory, Algorithms, and Applications*Siam.org; pp. 67-106.
37. Boriah S, V Chandola, V Kumar: Similarity Measures for Categorical Data: A Comparative Evaluation. *red* 2008; 30: 3.

38. Choi S-S, S-H Cha, C Tappert: A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics and Informatics* 2010; 8: 43-8.
39. Mead A: Review of the Development of Multidimensional Scaling Methods. *The Statistician* 1992: 41: 27-39.
40. Young FW, CH Null: Multidimensional Scaling of Nominal Data: The Recovery of Metric Information with Alscal. *Psychometrika* 1978; 43: 367-79.
41. Cox TF, MA Cox: *Multidimensional Scaling. Number 59 in Monographs on Statistics and Applied Probability*. Chapman & Hall 1994.
42. Borg I: *Modern Multidimensional Scaling: Theory and Applications*Springer, 2005.
43. Kristensen TS, H Hannerz, A Høgh, V Borg: The Copenhagen Psychosocial Questionnaire-a Tool for the Assessment and Improvement of the Psychosocial Work Environment. *Scandinavian journal of work, environment & health* 2005: 438-49.
44. Adibi P, AH Keshteli, A Esmaillzadeh, H Afshar, H Roohafza, H Bagherian-Sararoudi, et al.: The Study on the Epidemiology of Psychological, Alimentary Health and Nutrition (Sepahan): Overview of Methodology. *J Res Med Sci* 2012; 17: S291-7.
45. Alderman MH, H Cohen, S Madhavan: Dietary Sodium Intake and Mortality: The National Health and Nutrition Examination Survey (Nhanes I). *The Lancet* 1998; 351: 781-5.
46. Bruce B, JF Fries: The Stanford Health Assessment Questionnaire: A Review of Its History, Issues, Progress, and Documentation. *The Journal of rheumatology* 2003; 30: 167-78.
47. Frank E, M Hall: *A Simple Approach to Ordinal Classification* Springer, 2001.
48. Labovitz S: The Assignment of Numbers to Rank Order Categories. *American Sociological Review* 1970; 35: 515-24.
49. O'Brien RM: The Use of Pearson's with Ordinal Data. *American Sociological Review* 1979; 44: 851-7.
50. Mayer L: A Note on Treating Ordinal Data as Interval Data. *American Sociological Review* 1971; 36: 519-20.
51. Allen MP: Conventional and Optimal Interval Scores for Ordinal Variables. *Sociological Methods & Research* 1976; 4: 475-94.
52. Granberg-Rademacker JS: An Algorithm for Converting Ordinal Scale Measurement Data to Interval/Ratio Scale. *Educational and Psychological Measurement* 2010; 70: 74-90.
53. Ester M, H-P Kriegel, J Sander, X Xu: *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. in *KDD* 1996; pp. 226-31.
54. Ankerst M, MM Breunig, H-P Kriegel, J Sander: Optics: Ordering Points to Identify the Clustering Structure. *ACM Sigmod Record* 1999; 28: 49-60.
55. MacQueen J: *Some Methods for Classification and Analysis of Multivariate Observations*. in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*California, USA 1967; p. 14.
56. Kaufman L, P Rousseeuw: Clustering by Means of Medoids. 1987.
57. Bezdek JC: *Pattern Recognition with Fuzzy Objective Function Algorithms*Kluwer Academic Publishers, 1981.
58. Ball GH, DJ Hall: *Isodata, a Novel Method of Data Analysis and Pattern Classification*. DTIC Document 1965.
59. Defays D: An Efficient Algorithm for a Complete Link Method. *The Computer Journal* 1977; 20: 364-6.
60. Sibson R: Slink: An Optimally Efficient Algorithm for the Single-Link Cluster Method. *The Computer Journal* 1973; 16: 30-4.
61. Kaufman L, PJ Rousseeuw: *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344Wiley. com, 2009.
62. Xu R, D Wunsch: Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions on* 2005; 16: 645-78.
63. Sheikholeslami G, S Chatterjee, A Zhang: Wavecluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases. *The VLDB Journal* 2000; 8: 289-304.
64. Barbará D, P Chen: *Using the Fractal Dimension to Cluster Datasets*. in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*ACM 2000; pp. 260-4.
65. Andritsos P, P Tsaparas, RJ Miller, KC Sevcik: *Limbo: Scalable Clustering of Categorical Data*, in *Advances in Database Technology-Edbt 2004*Springer, 2004); pp. 123-46.
66. Barbará D, Y Li, J Couto: *Coolcat: An Entropy-Based Algorithm for Categorical Clustering*. in *Proceedings of the eleventh international conference on Information and knowledge management*ACM 2002; pp. 582-9.
67. Ganti V, J Gehrke, R Ramakrishnan: *Cactus—Clustering Categorical Data Using Summaries*. in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*ACM 1999; pp. 73-83.
68. Guha S, R Rastogi, K Shim: Rock: A Robust Clustering Algorithm for Categorical Attributes. *Information systems* 2000; 25: 345-66.
69. Parmar D, T Wu, J Blackhurst: Mmr: An Algorithm for Clustering Categorical Data Using Rough Set Theory. *Data & Knowledge Engineering* 2007; 63: 879-93.
70. Zaki MJ, M Peters: *Clicks: Mining Subspace Clusters in Categorical Data Via K-Partite Maximal Cliques*. in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*IEEE 2005; pp. 355-6.
71. Zhang P, X Wang, PX-K Song: Clustering Categorical Data Based on Distance Vectors. *Journal of the American Statistical Association* 2006; 101: 355-67.
72. Stutz J, P Cheeseman: *Autoclass — a Bayesian Approach to Classification*, in *Maximum Entropy and Bayesian Methods*Springer, 1996); pp. 117-26.
73. Huang Z: Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 1998; 2: 283-304.
74. Huang Z, MK Ng: A Fuzzy K-Modes Algorithm for Clustering Categorical Data. *Fuzzy Systems, IEEE Transactions on* 1999; 7: 446-52.
75. Kim D-W, KH Lee, D Lee: Fuzzy Clustering of Categorical Data Using Fuzzy Centroids. *Pattern recognition letters* 2004; 25: 1263-71.
76. Gan G, J Wu, Z Yang: A Genetic Fuzzy -Modes Algorithm for Clustering Categorical Data. *Expert Systems with Applications* 2009; 36: 1615-20.
77. Ng MK, MJ Li, JZ Huang, Z He: On the Impact of Dissimilarity Measure in K-Modes Clustering Algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2007; 29: 503-7.
78. Cao F, J Liang, D Li, L Bai, C Dang: A Dissimilarity Measure for the K-Modes Clustering Algorithm. *Knowledge-Based Systems* 2012; 26: 120-7.
79. Bai L, J Liang, C Dang, F Cao: The Impact of Cluster Representatives on the Convergence of the K-Modes Type Clustering. 2012.
80. Geweniger T, D Zülke, B Hammer, T Villmann: Median Fuzzy C-Means for Clustering Dissimilarity Data. *Neurocomputing* 2010; 73: 1109-16.
81. Brouwer RK, A Groenwold: Modified Fuzzy C-Means for Ordinal Valued Attributes with Particle Swarm for Optimization. *Fuzzy sets and systems* 2010; 161: 1774-89.
82. Johnson VE, JH Albert: *Ordinal Data Modeling*Springer, 1999.
83. Lee S-Y: *Handbook of Latent Variable and Related Models*Access Online via Elsevier, 2011.
84. Qu Y, MR Piedmonte, SV Medendorp: Latent Variable Models for Clustered Ordinal Data. *Biometrics* 1995: 268-75.
85. McParland D, I Gormley: *Clustering Ordinal Data Via Latent Variable Models*, in *Algorithms from and for Nature and Life*, ed. by Lausen B, Van den Poel D , Ultsch ASpringer International Publishing, 2013); pp. 127-35.
86. Lazarsfeld PF, NW Henry: *Latent Structure Analysis* New York,: Houghton, 1968; pp. ix, 294 p.

87. Bartholomew DJ, M Knott, I Moustaki: *Latent Variable Models and Factor Analysis: A Unified Approach*. Vol. 899Wiley. com, 2011.
88. Everitt BS: A Finite Mixture Model for the Clustering of Mixed-Mode Data. *Statistics & probability letters* 1988; 6: 305-9.
89. Everitt BS, C Merette: The Clustering of Mixed-Mode Data: A Comparison of Possible Approaches. *Journal of Applied Statistics* 1990; 17: 283-97.
90. Moustaki I: A Latent Trait and a Latent Class Model for Mixed Observed Variables. *British journal of mathematical and statistical psychology* 1996; 49: 313-34.
91. Magidson J, JK Vermunt: Latent Class Factor and Cluster Models, Bi-Plots, and Related Graphical Displays. *Sociological methodology* 2001; 31: 223-64.
92. McLachlan GJ, KE Basford: Mixture Models. Inference and Applications to Clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988* 1988; 1.
93. Vermunt JK, J Magidson: Latent Class Cluster Analysis. *Applied latent class analysis* 2002: 89-106.
94. Magidson J, JK Vermunt: Latent Class Models. *The Sage handbook of quantitative methodology for the social sciences* 2004: 175-98.
95. McLachlan G, D Peel: *Finite Mixture Models*Wiley. com, 2004.
96. Fraley C, AE Raftery: Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* 2002; 97: 611-31.
97. Moustaki I, I Papageorgiou: Latent Class Models for Mixed Variables with Applications in Archaeometry. *Computational statistics & data analysis* 2005; 48: 659-75.
98. Berzofsky M, PP Biemer: Weak Identifiability in Latent Class Analysis.
99. Lanza ST, LM Collins, DR Lemmon, JL Schafer: Proc Lca: A Sas Procedure for Latent Class Analysis. *Structural Equation Modeling* 2007; 14: 671-94.
100. Fraley C, AE Raftery: How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal* 1998; 41: 578-88.
101. Fraley C, AE Raftery: Mclust: Software for Model-Based Cluster Analysis. *Journal of Classification* 1999; 16: 297-306.
102. Muthen LK, L Muthén: Mplus [Computer Software]. *Los Angeles, CA: Muthén & Muthén* 1998.
103. Linzer DA, JB Lewis: Polca: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software* 2011; 42: 1-29.
104. Vermunt JK, J Magidson: Technical Guide for Latent Gold 4.0: Basic and Advanced. *Belmont (Mass.): Statistical Innovations Inc* 2005.
105. Haughton D, P Legrand, S Woolford: Review of Three Latent Class Cluster Analysis Packages: Latent Gold, Polca, and Mclust. *The American Statistician* 2009; 63: 81-91.
106. Ng MK, JC Wong: Clustering Categorical Data Sets Using Tabu Search Techniques. *Pattern Recognition* 2002; 35: 2783-90.
107. Morlini I: A Latent Variables Approach for Clustering Mixed Binary and Continuous Variables within a Gaussian Mixture Model. *Advances in Data Analysis and Classification* 2012; 6: 5-28.
108. Shih M-Y, J-W Jheng, L-F Lai: A Two-Step Method for Clustering Mixed Categroical and Numeric Data. *Tamkang Journal of Science and Engineering* 2010; 13: 11-9.
109. Huang Z: *Clustering Large Data Sets with Mixed Numeric and Categorical Values*. in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*Singapore 1997; pp. 21-34.
110. Ahmad A, L Dey: A< I> K</I>-Mean Clustering Algorithm for Mixed Numeric and Categorical Data. *Data & Knowledge Engineering* 2007; 63: 503-27.
111. Hsu C-C, C-L Chen, Y-W Su: Hierarchical Clustering of Mixed Data Based on Distance Hierarchy. *Information Sciences* 2007; 177: 4474-92.
112. Fayyad U, PS Bradley, CA Reina: *Scalable System for Clustering of Large Databases Having Mixed Data Attributes*. Google Patents 2003.
113. Mangasarian OL, WN Street, WH Wolberg: Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research* 1995; 43: 570-7.
114. Wolberg WH, OL Mangasarian: Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proceedings of the National Academy of Sciences* 1990; 87: 9193-6.
115. Mangasarian OL, R Setiono, W Wolberg: Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis. *Large-scale numerical optimization* 1990: 22-31.
116. Bennett KP, OL Mangasarian: Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimization methods and software* 1992; 1: 23-34.
117. Akay MF: Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis. *Expert Systems with Applications* 2009; 36: 3240-7.
118. Setiono R, H Liu: Neural-Network Feature Selector. *Neural Networks, IEEE Transactions on* 1997; 8: 654-62.
119. Xie XL, G Beni: A Validity Measure for Fuzzy Clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 1991; 13: 841-7.
120. Halkidi M, Y Batistakis, M Vazirgiannis: Clustering Validity Checking Methods: Part Ii. *ACM Sigmod Record* 2002; 31: 19-27.
121. Marateb HR, S Muceli, KC McGill, R Merletti, D Farina: Robust Decomposition of Single-Channel Intramuscular Emg Signals at Low Force Levels. *Journal of Neural Engineering* 2011; 8: 066015.
122. Goudarzi S, 'Clustering Ordinal Data : Diagnosing Functional Gastrointestinal Disorders - in Farsi', the University of Isfahan, 2013), p. 92.
123. Al-Riyami A: How to Prepare a Research Proposal. *Oman Med J* 2008; 23: 66-9.
124. Fathalla MF, MM Fathalla: *A Practical Guide for Health Researchers*World Health Organization, Regional Office for the Eastern Mediterranean, 2004.
125. Halkidi M, Y Batistakis, M Vazirgiannis: Cluster Validity Methods: Part I. *ACM Sigmod Record* 2002; 31: 40-5.

Source of Support: This study was supported by the University of Isfahan and Isfahan University of Medical Sciences, **Conflict of Interest:** None declared.