Copyediting 2: 15-20% revision

Children ~~, especially at early childhood ages (0 to 6 years),~~ are ~~more~~ susceptible to diseases, especially newborns and any one in early childhood with the age ranging from zero to six years old [1, 2]. ~~If~~ Treatment of ~~the~~ diseases in early childhood ~~is~~ is particularly important because~~not immediately treated, it can cause death~~mortality is at risk. Early c~~C~~hildhood~~'s~~ mortality rate each year is about 12.4 million [3]. ~~The~~ S~~s~~ymptoms that often ~~accompany~~ appear in early childhood~~'s~~ diseases are fever, cough, and diarrhea [4, 5, 6, 7, and 8].

Identification of early childhood diseases at an early stage is ~~necessary~~ extremely crucial for administering~~-~~ proper treatment and s~~r~~timulating recovery, but d~~. D~~iagnostic systems for early childhood diseases are often time-consuming and prone to errors. Moreover, ~~the~~ clinical decisions pertaining to disease diagnosis are mostly based on the intuition and experience of medical experts rather than the ~~knowledgeable~~ wealth of empirical data hidden in databases. ~~Therefore, T~~the former ~~ere~~often poses ~~is~~ a ~~possibility~~ risk of mis~~wrong~~diagnosis and mistreatment. Also, p~~P~~atients are usually advised to take a number of tests, ~~for the~~which are ~~disease diagnosis. In most of the case, not all the tests contribute to the effective~~ often inefficient or unnecessary in diagnosing~~s of~~ diseases.

Health care systems generate a huge amount~~s~~ of data containing hidden knowledge that is ~~cannot be discovered by~~inaccessible by traditional methods. The adaptable d~~D~~ata mining that ~~is~~ can be ~~a more adaptive technique for~~used in medical studies [9], is therefore one workable solution. Data mining ~~means searching for~~allows valuable information searching ~~-~~in large volumes of data, using automatic or semi-automatic exploration and analysis. Large quantities of data are explored and analyzed in order to discover meaningful patterns and rules.

Descriptive data mining is one of the major data mining types that will ~~. The descriptive methods~~ find ~~the~~ human interpretable pattern ~~that describes~~from ~~the~~ a messy set of data, whose process involve ~~: these include~~ clustering, association rule discovery and sequential pattern discovery. Meanwhile, a~~A~~ssociation rule mining is a ~~the~~ data mining process used to find ~~the~~ rules that may govern associations and causal objects between item sets. As for Apriori algorithm, it is a method used to find ~~the~~ relationship patterns between one or more items in a dataset. Apriori association technique has been proven to be effective in finding various trends in healthcare databases [10]. ~~Apriori algorithm~~It is well known ~~as one of the most representative algorithms fo~~for its representativeness in~~r~~ data mining [11, 12] and it ~~. This algorithm~~ deals with ~~the~~ item sets called transactional data [13]. Considering the relevancy of association rule mining in ~~disease diagnosis and~~ the ~~issue of dynamic update of~~rapidly changing medical databases, and thus in disease diagnosis, this paper ~~presents~~ attempts to provide a more~~n~~ efficient technique to identify ~~the~~ risk factors for early childhood diseases by mining association rules from the dynamically changing medical databases.

~~The medical record of the patient~~A patient medical record generally consists of ~~various~~ a myriad of features, so ~~. Therefore,~~ feature reduction is very important to identify the most significant risk factors related to each disease [14]. Feature reduction has been an active and fruitful field of research in machine learning and data mining. Feature reduction is a dimensionality reduction technique used to reduce irrelevant data and to increase accuracy [15]. Principal Component Analysis (PCA) is one ~~of the well-known~~popular statistical technique~~s~~ aiming~~s~~ to reduce data dimensions without losing any important information ~~from data~~ [16, 17]. PCA basically converts and decomposes a large number of uncorrelated variables into a smaller number of correlated variables and ~~can reduce~~deductible data dimensions [18]. PCA has several advantages such as reducing data redundancy, ~~reducing~~ complexity, ~~reducing~~ database size, and ~~reducing~~ noise and it ~~. PCA~~ can be used to discover ~~the~~ correlation between variables [18].

This study ~~provides~~ incorporates feature~~s~~ reduction technique and association rule mining technique. Apriori algorithm was used to generate the pattern sets. PCA ~~Features reduction technique~~ was used as a feature reduction technique to generate the factors contributing to e~~E~~arly c~~C~~hildhood diseases. As for the association rule mining, we developed ~~. PCA was used for feature reduction technique. a~~An efficient pattern mining technique ~~has been developed~~ that ~~. It~~ can proficiently derive

new set of patterns and rare rules from updated databases with faster execution time and less space usage, without any loss of information. To the best of our knowledge, ~~it~~ this is the first attempt to generate ~~the~~ a complete set of association rules from dynamically changing medical databases ~~for~~ to identify~~ing~~ risk factors for early childhood diseases. ~~Apriori algorithm was used to generate the set of patterns.~~ To summarize, the major contribution~~s~~ of this paper ~~are~~ is ~~an~~ the identification of factors contributing to early childhood diseases ~~using the proposed approach,~~ and the efficient generation of the set~~s~~ of patterns and association rules with updated threshold values ~~using the proposed approach~~, a~~.~~ll while designing a new proposed approach that facilitated these two processes.

~~This paper is organized as follows. the introduction is presented in the first section, then~~In what follows, the methods ~~used for~~of feature~~s~~ reduction and association rule mining are presented ~~in the next section~~, continued with the description of the research result~~s~~ and discussion ~~in the third section~~, ~~and finally, the~~which is then followed by the inferred conclusion~~s~~ ~~is presented in the last section~~.

[…]

In general, the PCA technique transforms n vectors $(x_1, x_2, \ldots\ldots, x_i, \ldots . x_n)$ from a-d dimensional space to n vectors $(x'_1, x'_2, \ldots\ldots, x'_i, \ldots . x'_n)$ in a new, d -dimensional space as [19]

(1)

where $\boldsymbol{e}_k$ are the eigenvectors corresponding to the d' largest eigenvalues for the *scatter matrix* **S**. *The principal components* of the original data set are the projections of the original vectors x_i on the eigenvectors $\boldsymbol{e}_k$, which is denoted $a_{k,i}$. Both d and d' are positive integers, and the dimension d' cannot be greater than d . The d -by d - scatter matrix **S** for the original data set $(x_1, x_2, \ldots\ldots, x_i, \ldots . x_n)$ is defined as:

(2)

where $\mathbf{E}\left[x_i x_i^T\right]$ is the statistical expectation operator applied on the outer product x_i of and its transpose. The representation shown in (1) minimizes the error between the original and transformed vectors. This is illustrated by considering the variance of the -principal components given by [23]:

(3)

where $\boldsymbol{e}_k$ represents the d -by-1 vector $\boldsymbol{e}_k = \left[\boldsymbol{e}_{1,k}\boldsymbol{e}_{2,k} \ldots \boldsymbol{e}_{d,k}\right]^T$.

It is evident that the variance of the principal components is a function of the magnitude of the components of the vectors $\boldsymbol{e}_k$. At the local maxima and minima for the variance function in (3), the following relationship exists:

(4)

(5)

(6)

Equation can be ~~recognized as~~considered an eigenvalue problem with nontrivial solutions only when $\boldsymbol{\lambda}$ is the eigenvalue~~s~~ of the scatter matrix. Thus, the associated vectors $\boldsymbol{e}_k$ (k = 1 to d) are the eigenvectors. If the condition d `< d is satisfied, then the above representation also reduces the dimensionality of the vectors. The error in representation of the original data set $(x_1, x_2, \ldots\ldots, x_i, \ldots . x_n)$ due to the reduction in the number of dimensions to is given by [23]

Wordcount: 995
Revisions: 200 (100 insertions and 100 deletions)