Osteoporosis Identification Using Data Mining Techniques

♦ Recepción/ 27 junio 2019
♦ Aceptación/ 25 agosto 2019

Osteoporosis Identification Using Data Mining Techniques

Identificación de osteoporosis utilizando técnicas de minería de datos

Eman Turky[1], Raghad Hameed Ali Al-abboodi[2], Wijdan Jaber AL-kubaisy[3], Maha Mahmood[4]

[1]College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq. maymoonat@gmail.com
[2]Business Informatics College, UOITC, Iraq. raghad.hav@uoitc.edu.iq.
[3] College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq, wijdan@computer-college.org2
[4] College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq. maha_mahmood@computer-college.org

ABSTRACT/ Data Mining is a technique for discovering information results from large databases. A large database represents a huge amount of information that can be potentially very useful if discovered and summarized correctly. This paper presents a research in developing data mining ensembles for predicting the risk of osteoporosis prevalence in human. Osteoporosis is a bone disease that commonly occurs among postmenopausal women and no effective treatments are available at the moment, except prevention, which requires early diagnosis. However, early detection of the disease is very difficult. This research aims to devise an intelligent diagnosis support system by using data mining ensemble technology to assist General Practitioners assessing patient's risk at developing osteoporosis. This paper describes the methods for constructing effective ensembles through measuring diversity between individual predictors. Apriori-PT are implemented by neural networks training. The ensembles built for predicting osteoporosis are evaluated by the real-world data and the results indicate that the algorithm has relatively high-level of diversity and thus are able to improve prediction accuracy.
Keywords: Data Mining, Osteoporosis, Apriori-PT, Exploratory Data Analysis, Knowledge Discovery in Databases, Neural Network.
RESUMEN / **La minería de datos es una técnica para descubrir resultados de información de grandes bases de datos. Una gran base de datos representa una gran cantidad de información que puede ser potencialmente muy útil si se descubre y resume correctamente. Este artículo presenta una investigación en el desarrollo de conjuntos de minería de datos para predecir el riesgo de prevalencia de osteoporosis en humanos. La osteoporosis es una enfermedad ósea que ocurre comúnmente entre las mujeres posmenopáusicas y no hay tratamientos efectivos disponibles en este momento, excepto la prevención, que requiere un diagnóstico temprano. Sin embargo, la detección temprana de la enfermedad es muy difícil. Esta investigación tiene como objetivo idear un sistema inteligente de soporte de diagnóstico mediante el uso de la tecnología de conjunto de minería de datos para ayudar a los médicos generales a evaluar el riesgo del paciente en el desarrollo de osteoporosis. Este artículo describe los métodos para construir conjuntos efectivos a través de la medición de la diversidad entre predictores individuales. Apriori-PT se implementan mediante capacitación en redes neuronales. Los conjuntos construidos para predecir la osteoporosis son evaluados por los datos del mundo real y los resultados indican que el algoritmo tiene un nivel de diversidad relativamente alto y, por lo tanto, puede mejorar la precisión de la predicción.**
Palabras clave: minería de datos, osteoporosis, Apriori-PT, análisis exploratorio de datos, descubrimiento de conocimiento en bases de datos, red neuronal.

1. INTRODUCTION

In the daily life of people, they usually come across a situation like: a message telling them that a store or an e-commerce platform is presenting new merchandises or promotions, or e-mails notifying them about an upcoming celebration of a mall anniversary. This is due to the fact that store retailers check the data of the customers and gather that data, retailers can benefit from this data, in addition to the e-commerce records, shopping log, etc., a more extensive understanding of the needs of consumers for the sake of making an analysis of correlation. In the movement of big data age, tools of data analysis in the systems of information management haven't obtained data from massive complicated data. For instance, information systems cannot analyze data which have been obtained and observed from the surface of the satellite, Marine atmospheric, due to the properties and size of space and time. Moreover, there are huge genetic data amounts. Thereby, there is a need for developing a new approach for data mining. It is now largely, however not universally understood that the precision of the training set, derived with the use of the training data for testing as well, may be highly optimistic. Cross-validation or a bootstrap method is thus more preferable. Where model tuning and/or feature selection are an element of the process of model fitting, care is needed for avoiding subtler versions of the bias in the accuracy measure of the training set. For an unbiased evaluation, any model tuning and/or feature selection should be repeated at every one of the cross-validation folds [14].

2. DATA MINING

Data mining has been around for many years. The term becomes Big Data mining when mining involves huge amount of data with characteristics such as volume, velocity and variety. Big data mining assumes importance as the enterprises are producing data with exponential growth. Big data mining refers to mining voluminous data and extracting comprehensive business intelligence. Therefore, it provides opportunities for enterprise to leverage Big Data mining for making expert decisions. At the same time, security is an important concern. From the literatures, it is understood that Big Data mining has both opportunities and security implications. Two important facts were conceived from review of literatures. First, "Big data mining provides more opportunities for business growth and transformation". Second, "Big data mining can have security implications". Data Mining is the advanced procedures that extract the possible, efficient and understandable mode from the huge data amounts in accordance with the established business goals [1]. Many people consider data mining as common term of knowledge discovery, whereas others merely put mining as a main step in the knowledge discovery procedure [2]. It emerged in the late 80s, and it is an area with a considerable research value in the study of the data-base and overlapping subject, in combination with AI, pattern identification, data-base technology, statistics, machine learning, data visualization and other areas of technology and theory. As a technology type, the lifecycle of data mining is in vague stage, specialists need to take energy and time for researching, enhance and boost up to gradually mature, and ultimately be accepted [3,4]. Data mining is a technology type that combines the conventional approaches of data processing with various methods, for the analysis of new types of data and for extracting knowledge from massive data amounts. There are two kinds of knowledge that can be found in huge amounts of data, one is online analytical processing (OLAP), the other one is a data mining (DM). each of those is analysis tools based on data warehouse, but on-line analytical processing appears earlier than the DM, according to a multi-dimensional view, focusing on executing efficient and fast response for the commands of the user; DM pays attention to the beneficial model to individuals that are hiding in the depths of data and it is performed via automation, with no customer participation.

2.1 DATA MINING PROCESS PHASES

There are number of stages in a DM and Knowledge Discovery in Databases DM/KDD process that have no much different from the process of any operations research project or comprehensive software engineering [9,10], these are: goal determination (20% of total time), preparation of the data (60% of total time), mining the data (10% of total time), analysis of results (10% of total time) and assimilation of the extracted knowledge [11,12]. Figure 1 illustrates those phases [9].

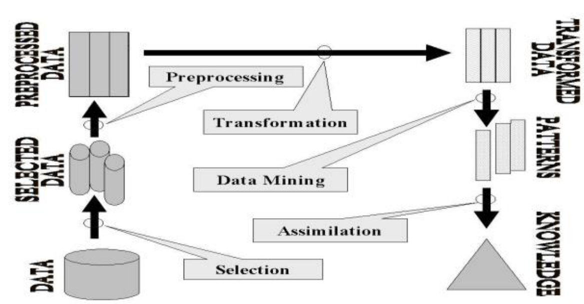

Figure 1: Data Mining Process Phases.

Data Selection is, in fact, a complicated *sub-task*. Initially, it includes the definition of the variables providing the data and specifying the suitable sources of data. After that, (construction of meta-data), requires understanding and defining every data element like data types, possible values, formats, etc. Then, need to retrieve the data, which is not always straightforward. For instance, could need to search the web or a data warehouse. Internet searches, redundant in applications of DM/KDD could generate many of matches, a great deal of which are not relevant to the query [10, 11].

In a context like that, issue of storing and retrieving information must be taken under consideration with a great deal of carefulness. One more related issue of information management is the impact of context (data model) in managing knowledge that might be characterized as the aggregation of data with context, for a certain objective. As a result, the significance of the analysis of the design of the data-base and matters of usage, as partly preparing to the data phase [10,11].

The ***task*** *of* ***data preprocessing*** is involved with making sure that the quality of the chosen data is adequate. This is highly important in the case where the data which will be mined has many different sources, like numerous organization departments. This phase is involved with, for example, the removal of inconsistencies in names of attribute or their value names between datasets of a variety of sources [13].

A ***sub-task*** *of* ***data transformation*** could as well be required in the case where various data come in units that are not compatible with one another. To recapitulate, the discussion above focuses on the importance of the implementation of a thorough and well planned control of data quality and assurance step, before the step of data analysis. In addition, they underline needing to improve the processes of data collection. All those time-consuming actions are crucially important in the process of DM/KDD [10,11].

3. OSTEOPOROSIS

Osteoporosis is an illness which result in reducing the density and quality of bones, which in turn results in weakening the skeleton and increasing risk of bones getting fractured. This is a universal issue that affects human particularly, 1/3 of humans get osteoporosis throughout their life [7-8]. In people that have this disease, there is a gradual bone loss and with no apparent signs to the point where the illness advances to its late phase. This is why, this illness is usually called the "silent epidemic". The diagnosis of this illness is typically done via conducting tests with the use of a type of scanning tools, such as Peripheral Dual-Energy X-ray Absorptiometry (PIXI), Dual-Energy X-ray Absorptiometry (DEXA), or Quantitative Ultrasound (QUS). Those tests evaluate bone mineral density (BMD) in a variety of ways or at various places. DEXA is the golden standard; on the other hand, having access to such type of equipment is still not adequate even in the majority of developed countries. The key motivation of the present study is investigating substitute computerised, intelligent ways of aiding to early diagnose Osteoporosis [10].

4- Related work

There are some other works which investigated this path of research, Ordonez et al. [16] for example determined the osteoporosis in postmenopausal women. At their work, they applied a machine learning techniques by taking a dataset of the relationship between BMD, diet and lifestyle habits of 305 postmenopausa women, and by using the regression support vector to yield a non-linear model. A mathematical model is resulted by sing the SVMs which identify the relationship, while a regression tree is applied to the output SVMs to generate the greatest weight in the relationship. As a result, they found out few reasons such as the extra taking for the calcium, exposing to the sun in a suitable level, weight control is all consider main factors to control this illness. In Moudani et al. [17] the authors suggested a predictive system which it could detect the early stages of the osteoporosis in adults. At their paper, a data has been collected by interviewing many patients using a tool called FRAX tool which provided a probability of this illness for 10 years. As a result, they proved that age, body mass index, alcohol and smoking are all represent relation to osteoporosis risk. Their model was based on decision tree and and random forest algorithms where they proved that random forest algorithm generate the highest performance than the decision trees algorithm.

using wrapper-based feature selection, Hsueh-Wei et al. [18] provided a comparisons of three different classification algorithms. The three algorithms were multilayer feed-forward neural network (MFNN), Naïve Bayes and logistic regression. The output of their research showed that the MFNN outperformed the other algorithms when predicting the disease. This study suggested the method for the doctors and the patient to improve making decision related to the illness.

However, in our study the neural network and decision trees combined, have been implemented for the prediction of prevalence risk of human osteoporosis. They proved better accuracy and detection through several of training and testing phases.

5. DATA SET

Bio-GPS (http://biogps.org) is a centralized geneannotation portal which gives researchers the ability of accessing resources of distributed gene annotation. This paper is focused on the updates to Bio-GPS due to the fact that the last paper (2013 data-base issue). The distinct Bio-GPS characteristics, in comparison with the ones of other portals of genes, are its user customizability and community extensibility. Users take part in the gene-specific resources that are accessible from Bio-GPS ("plugins") that is helpful in making sure that the collection of resources is constantly updated and that it'll keep expanding time after time. Users of Bio-GPS can generate their own sets of relevant plugins and save those plugins as customized pages of gene reports or "layouts".

For all registered plugins, the "Gene expression/activity chart" plugin (or 'data chart'; http://biogps.org/ plugin/9/) is the most widely used. And it is as well one of a few plugins that are maintained and developed by the Bio-GPS team. The "data chart" plugin is included in more than half of user created layouts (1408 / 2688). Based on its popularity, the preliminary release of the 'data chart' came with many gene-expression reference data-sets. Additionally expanded the set via including about 6000 data-sets from the Array Express repository of EBI. Those data-sets come from 9 common micro-array platforms from humans, rats and mice. Loaded those meta-data to Bio-GPS in order to give users the ability of customizing the visualization of expression profiles on the basis of the sample annotations. The figure2 show the example of the data set we are used [14].

Factors: (via ArrayExpress)

Sample	SAMPLE ID	GENDER	AGE	PASSAGE
GSM878108	donor 5 with primary osteoporosis (MSC population #573)	female	89 yrs (elderly)	passage 1
GSM878107	donor 4 with primary osteoporosis (MSC population #572)	female	82 yrs (elderly)	passage 1
GSM878106	donor 3 with primary osteoporosis (MSC population #558)	female	87 yrs (elderly)	passage 2
GSM878105	donor 2 with primary osteoporosis (MSC population #547)	female	94 yrs (elderly)	passage 1
GSM878104	donor 1 with primary osteoporosis (MSC population #535)	female	79 yrs (elderly)	passage 1
GSM878103	donor 4 (MSC population #663)	female	89 yrs (elderly-aged)	passage 1
GSM878102	donor 3 (MSC population #606)	female	80 yrs (elderly-aged)	passage 1
GSM878100	donor 1 (MSC population #520)	female	79 yrs (elderly-aged)	passage 1
GSM87810	donor 2 (MSC population #559)	male	79 yrs (elderly-aged)	passage 1

Figure2: Sample of The Data Set.

6. TRAINING PHASE

Building NNs is possibly the most widely used practice in the community of DM simply because of its low cost and the fact that it requires less effort in the training of NNs. The related researches show that the efficiency of this type of groups differs and some of them are not so good. This is due to the fact that the models, (trained NNs) are very correlated and thereby have the tendency of making the same mistakes in a simultaneous manner. This is why, attentions must be paid for generating more varying NNs to build sufficient sets via measuring the difference amongst the trained networks. Multiple Layer Perceptron (MLP) NNs have been trained as the potential models for the sets. [13], in addition to the approach of random sampling, for training and therefore, the NNs have been trained for more diversity. The process which is described below has been utilized to build an NN ensemble (NNE). For every sub-set of the chosen characteristics: a) Ten-fold cross validation approach has been utilized to train and validate every network. Moreover, three NNs with various set-ups in the hidden layer have been trained and implemented for every partition of the data. Therefore, a total of 100 networks have been generated and put in a pool as the candidates of the ensemble member models b) splitting the dataset into training and testing sets c) Building ensembles: some of the networks that have higher level of diversity are selected to construct ensembles together with some of them randomly selected from the candidate pool d) Calculating the diversity amongst the networks e) Assessing the efficiency with the testing dataset.

The process of developing a model that understands a data source. In neural networks, the process of adjusting the connection weights in a neural network under the control of a learning algorithm.

7. MATERILAS AND METHOD

7.1 Apriori-PT

Apriori-PT (Apriori on Parts of Transactions of database) is Apriori-based association rule mining algorithm. In fact it is for first step for discovering association rule that interested with finding all possible frequent itemsets. It is similar to Apriori algorithm in progress, it uses the same procedure of candidate generating, but do not use all transactions of database to count the support of an interested itemset[5-6]. Rather it uses just the database transactions that have length equal or greater than the length of the interested itemset. Apriori-PT algorithm is shown below:

Table 1: Notations of Apriori-PT ARM Algorithm

Notation	Meaning
k-itemset	An item-set have *k* items
C_k	Gropus of Candidate k-item-sets (possibly large item-sets) Every member of this group has 2 fields: Item-set and Support count
t	Transactions of Database
$\|t\|$	Length of ***t***
R_k	Group of *k*-item-sets (the ones with minimum support) Every member of this group has 2 fields: Item-set and Support count.
c	Candidate

```
D = Database
R1 = all frequent used items
k = 1
Repeat
    Ck = apriori-generate (Rk-1)
    For each transactions t ∈ D where |t| ≥ k
        Increment Count for any c in Ck where c ⊆ t
    Rk = {c∈Ck where Count(c) ≥ minsup}
        Frequent_Itemsets = Frequent_Itemsets ∪ Rk
            ++k
Until (Rk is empty) or ( k = maximum length of transactions)
Return Frequent_Itemsets
```

Figure 3: Apriori-PT ARM Algorithm

8. RESULTS

DM can be considered as an optimized type of Exploratory Data Analysis (EDA). Because of their huge dimensions, data mining issues are iterative procedures. As soon as every iteration is done, its results are utilized in one of 2 ways: as the input for a new iteration (if required) or as a stage-final (in other words, final for the ongoing stage). That is, to the point where new data arrives and a re-analysis is needed.

Data mining is a team effort. Domain experts, analysts and other involved team members, work together and collectively process the results of the analysis. The collective approach offers a key reason for the iterative nature of data mining and one of its strongest properties. For instance, a few team members might not be pleased with the results of the analysis and might be wishing to improve them, or experimenting with other variations of the issue, according to this type of results. This situation re-initiates (and enriches) the entire process of data mining analysis.

An association rule mining algorithm is used based on the Apriori approach. This algorithm is referred as Apriori-PT (Apriori on Parts of Transactions of database).

Apriori-PT uses the basic idea and tools of standard Apriori algorithm such as apriori-generate procedure for generating the candidates, but the modification is happened on the technique of applying the support counting on the database.

Table2: The Performance of the A *Apriori-PT.*

Number of neural	Voting	Mean	Specificity	Sensitivity
100	72.4	68.4	0.723	0.678
200	71.3	69.5	0.745	0.655
300	70.2	64.5	0.737	0.687
400	69.9	65.6	0.765	0.632
500	72.8	67.4	0.778	0.67
600	65.6	65.6	0.728	0.675
700	75.7	78.5	0.824	0.772
800	87.87	85.6	0.865	0.731
900	82.8	87.1	0.888	0.776
1000	89.7	88.7	0.869	0.735

Table 2 summaries the performances test of the sets of size Number of neural network that vary from 100 up to 1000. For every Neural, 100-fold cross validations have been performed and after that, 100 sets have been created respectively. It is clear that on average these mixed sets performed better than other types.

Table 3: The Number of the Training and Testing.

Model	Train	Test
10	78.67	70.56
20	88.89	67.66
30	77.78	60.65
40	95.9	61.67
50	82.67	59.65
60	89.56	61.68
70	77.98	69.6
80	87.87	65.67
90	85.65	70.65
100	89.78	79.9
Mean	85.475	66.769
Specificity	0.886	0.843
Sensitivity	0.884	0.785

In table3 It is seen that 100 models created about 89.86% precision in training, but merely 79.90% on the testing data. On the other hand, due to the fact that these models are of quite a high degree of diversity of CFD, the group that has been built with them has been projected to produce more efficient results, coincident diversity of failure and, in fact, reached 85.475% and 80.29% with the use of the majority-vote, nearly 5% enhancements on the average precision of each of training and testing, an encouragement for creating more groups of different sizes and types, with the use of risk factors that have been chosen by the salience methods of estimation.

9. CONCLUSIONS

DM is an improved type of EDA. Because of their massive dimensions, data mining issues are iterative procedures. The used algorithm is referred as Apriori-PT (Apriori on Parts of Transactions of database). The ensemble types, in other words, ensembles of NN alone, decision trees alone, and combined, have been implemented for the prediction of prevalence risk of human osteoporosis. Their efficiency has been evaluated based on the coincident failure diversity (CFD), sensitivity, generalization accuracy, and specificity. The measurement of the diversity amongst models gives a useful knowledge about the groups and gives an explanation why some of the ensembles generate gains and some don't. For single type models, of either decision tree or NN, the widely utilized mechanisms of training have the tendency of generating considerably dependent models that have a small level of diversity amongst homogeneous models, for example either amongst the networks or the decision trees. This is why, the groups that have been built with homogeneous models won't generate much or any gain on the efficiency. However, hybrid groups which have been created with combined models of NNs and decision trees are of rather higher coincident-failure diversity and for this reason are capable of improving the precision. The results of experiments that have been presented in the present study indeed confirmed that the hybrid groups if created with considering diversity, performed better than other homogeneous groups in addition to the separate predictors which operated solely. On the other hand, the diversity amongst the trained NNs and the induced decision trees is still not sufficiently high for considerably improving the efficiency. Apriori-PT uses the basic idea and tools of standard Apriori algorithm. The support counting process is achieved by scanning part of transactions. The association rules generating phase are classified according to minimum confidence and using special mechanism for ignore the rules that not interested.

10. FUTURE WORKS

Various techniques of DM were researches for the sake of processing and analysing numerous variations of Osteoporosis disease, where the most commonly known tasks of DM are summarization, association rules mining, classification, and clustering. This is why, more researches must be addressed in the directions of analytics of big data for energy of wind for optimizing the designs of wind farms for the efficient prediction of the generated power. More research must be focused on researching the methods of enhancing the diversity amongst the models.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.

[2] A. Krogh, J.Vedelsby. "Neural network ensembles cross-validation and active learning". In Tesauro, Touretzky, and Leen, eds, Advances in Neural

Information Processing System. vol.7. pp231-238, MIT press.1995.

[3] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules". In Proc. of the 20th Int'l Conference on Very Large Databases (VLDB '94), Santiago, Chile, June 1994.

[4] Ogudo, K.A.; Muwawa Jean Nestor, D.; Ibrahim Khalaf, O.; Daei Kasmaei, H. A Device Performance and Data Analytics Concept for Smartphones' IoT Services and Machine-Type Communication in Cellular Networks. *Symmetry* **2019**, *11*, 593.

[5] D. partridge, W.B. Yates. "Multiversion neural network systems. Proceedings of neural networks and their applications", NEURAP'98, Marseilles, France, pp351-357. 1998.

[6] W. WangP. JonesD. Partridge. "Assessing the Impact of Input Features in a Feedforward Neural Network", Vol9, pp101-112, 2000.

[7] GHAIDA MUTTASHAR ABDULSAHIB and OSAMAH IBRAHIM KHALAF, 2018. AN IMPROVED ALGORITHM TO FIRE DETECTION IN FOREST BY USING WIRELESS SENSOR NETWORKS.International Journal of Civil Engineering & Technology (IJCIET) - Scopus Indexed.Volume:9,Issue:11,Pages:369-377.

[8] A. A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", Postgraduate Program in Computer Science, University of Catolica, 1155. Curitiba - PR. 80215-901. Brazil.2002.

[9] L. I. Kuncheva, Ch. J. Whitaker." Measures of diversity in classifier ensembles and their relationships with the ensemble accuracy". Kluwer Academic Publisher, Netherlands, 2003.

[10] R. K. Aaron, D. M. Ciombor, and B. J. Simon, "Treatment of nonunions with electric and electromagnetic fields," Clinical Orthopaedics and Related Research, no.419, pp.21–29,2004.

[11] Osamah Ibrahim Khalaf, Bayan Mahdi Sabbar''An overview on wireless sensor networks and finding optimal location of node'',Periodicals of Engineering and Natural Sciences, Vol 7, No 3 (2019)

[12] K. T. Lim, J. H. Kim, H. Seonwoo et al., "In vitro effects of electromagnetic field stimulation on cells in tissue engineering," Tissue Engineering and Regenerative Medicine, vol.6, no.4, pp. 675–684,2009.

[13] B. Al-Khateeb and M. Mahmood, "A Framework for an Automatic Generation of Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, 1694-0784, 2013.

[14] Ayman Dawood Salman1, Osamah Ibrahim Khalaf and Ghaida Muttashar Abdulsahib, 2019. An adaptive intelligent alarm system for wireless sensor network. Indonesian Journal of Electrical Engineering and Computer Science, Vol. 15, No. 1, July 2019, pp. 142~147

[15] M. Mahmood and B. Al-Khateeb, "Towards an Automatic Generation of Neural Networks", Journal of Theoretical and Applied Information Technology, Vol.95. Issue 23, 2017.

[16] Osamah Ibrahem Khalaf, Ghaidaa Muttasher et al., "Improving video Transmission Over Heterogeneous Network by Using ARQ and FEC Error Correction Algorithm", vol. 30, no.8, pp.24-27, Nov 2015

[17] W. Moudani, A. Shahin, F. Chakik, and D. Rajab," Intelligent Predictive Osteoporosis System", International Journal of Computer Applications 32 (5): 28–37.2011.

[18] Osamah Ibrahim Khalaf, Ghaida Muttashar Abdulsahib and Muayed Sadik, 2018. A Modified Algorithm for Improving Lifetime WSN. Journal of Engineering and Applied Sciences, 13: 9277-9282