ISSN: 0067-2904

Proposed Approach for Analysing General Hygiene Information Using Various Data Mining Algorithms

Tareef K. Mustafa*, Mustafa S. Abd
Department of Computers, College of Science, University of Baghdad, Baghdad, Iraq.

Abstract

General medical fields and computer science usually conjugate together to produce impressive results in both fields using applications, programs and algorithms provided by Data mining field. The present research's title contains the term hygiene which may be described as the principle of maintaining cleanliness of the external body. Whilst the environmental hygienic hazards can present themselves in various media shapes e.g. air, water, soil…etc. The influence they can exert on our health is very complex and may be modulated by our genetic makeup, psychological factors and by our perceptions of the risks that they present. Our main concern in this research is not to improve general health, rather than to propose a data mining approach that will eventually give a more clear understanding and automotive general steps that can be used by the data analyser to give more enhanced and improved results than using typical statistical tests and database queries. This research proposes a new approach involving 3 algorithms selected from data mining which are association rule mining, Apriori algorithm and Naïve Bayesian consequently, to offer a final improved decision support results that can serve the researchers in their fields.

Keywords: Data Mining, Association Rule, Apriori, Naïve Bayesian, Hygiene Information.

منهج مقترح لتحليل المعلومات الصحية العامة باستخدام خوارزميات تنقيب بيانات متعددة

طريف كامل مصطفى*، مصطفى سلمان عبد
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق.

الخلاصة

حقول الطب العامة وعلوم الحاسوب عادة ما تنتج معلومات جديرة بالاهتمام عند تسخيرهما معا بإستثمار التطبيقات، البرامجيات، والخوارزميات الخاصة بتنقيب البيانات. عنوان البحث الحالي يتضمن مفردة الصحة العامة والتي من الممكن توصيفها بانها مبادئ إدامة النظافة لعموم جسم الإنسان. بينما تتضمن مبادئ الحفاظ على البيئة تجنب مخاطر التلوث البيئي بأشكال ومظاهر عديدة مثل التربة، الماء، الهواء...الخ. إن التأثيرات التي تفرزها هذه العوامل البيئية على صحتنا العامة متشعبة وقد تؤثر علينا بعدة أشكال كأن تكون جينية، نفسية، أو اي شكل اخر إعتماداً على منظورنا البحثي. في هذا البحث لاينصب اهتمامنا البحثي في العناية

*Email: tareef_alshaibi@yahoo.com

بالصحة العامة بقدر ما يكون اهتمامنا باقتراح منهج تنقيب بيانات يصب بالنهاية في صالح البحثين الصحي والمعلوماتي، كما ويعطي صورة أكثر وضوحا مع خطوات الية مبسطة يتبعها الباحث الطبي والصحي حتى وإن لم يكن ضليعا بعلوم تنقيب البيانات، بحيث يعطي نتائج أكثر عمقا ودقة مما لو استخدم الادوات الاحصائية التقليدية الشائعة. البحث الحالي يقترح نهجا يتضمن دمج ثلا ث خوارزميات تعمل بالتعاقب على البيانات الصحية لتعطينا نتائج باهرة لم تكن ضمن خطة الباحث عند تصميمه تجاربه وتوفر له الية اتخاذ قرارات تخدم مجال بحثه.

Introduction

Researchers in general hygiene fields are using simple statistical analysis tools such as mean, variance, t-test...etc. to obtain and extract their results. Using such basic methods will probably provide primitive and simple conclusions. Although those professionals and researches maybe specialized in their own scientific fields, but they may be unaware of the tools, algorithms, and applications of data mining that can support their researches, their blurred overview in using these tools to serve the researcher objectives will lead to unnecessary time waste.

Due to the wide variations of data mining techniques and its capabilities for data analysis in various fields of research, this gave them a wide popularity for discovering information in a comfortable way after a clear understanding of the problem. In addition, the data mining techniques have the ability to extract hidden patterns and clarify relationships among attributes of data especially in medical fields. These are what the researchers need to know for supporting decision makers for detecting and diagnosing diseases in early stage [1].

The main objective for this research is to produce a clear and consistent approach that can be used by general health researchers and can guide them through directed steps to reach their objectives without losing time or efforts. Eventually, these steps in the proposed approach will produce better understanding to the facts and therefore will lead to various results that could not be extracted using the primitive and statistical common tools which was mentioned earlier.

Research specifications are appointed to describe the proposed approach area along with its limitations since they are both part and parcel from the main design. They are summarized as follows:

- A general hygiene questionnaire form was designed and distributed into 200 students in 2 high schools in Baghdad to be filled by the students themselves with guidance and observation done by the research team in addition to the help of the school management.
- The chosen schools were high schools for boys and girls within the age of 16-18 years to be able to fill the forms and have some partial dependency in their hygiene way of life, considering what they eat, their chosen transportation to attend school, their home environmental status…etc.
- The data represents general environmental health characteristics, and does not refer to the extremes of the climate occupational hazards.
- WEKA 3.7.10 was used for its excellence performance in such data mining implementation analysis.
- The proposed approach uses 3 data mining algorithms which are association rule mining, Apriori algorithm and Naïve Bayesian consequently to offer a final improved decision support result.

Literature review

Lots of work has been done in the general medical fields research to catch unknown new relations to improve the human been health in all ages. Here is a short review showing some examples for what has been done in this area using data mining techniques:

Bajaj, Choudhary and Chauhan [2] basically extended their work in detection of diseases using data mining techniques. The work showed that the heart disease can be diagnosed by using data mining techniques and algorithms such as decision tree and other applied models. Attributes used in their research such as age, smoking, root canal treatment, and the diabetes disease predicts the ratio of the probability for patients getting a heart disease. The system they proposed was the union of computer-based patient records using clinical decision support could lower down the medical mistakes, and help to improve patient's safety and their overall outcome.

Von, Huttunen, Vihavainen, et al. [3] presented their work on information extraction application technology to the domain of the Public Health in a real-world scenario. The research presented two

novel points, the first is to distinguish a quality criteria and its objective is to measure correctness of the designed system's analysis in traditional terms using F-measure, not forgetting the subjective criteria that measure the utility of the results to end-users. Second, is to obtain measures of utilities, building an environment that allows users to use an interacting system by rating the analysed content. Then building and comparing several classifiers to learn from the user's responses to mainly predict the relevance scores for the new events. They conducted experiments with learning to predict relevance to discuss the results and their implications in text mining health domain.

A research team containing Ahmed, Eshlaghy, Poorebrahimi, et al. [4] used several data mining techniques for developing models to predict the recurrence of Breast cancer using data collected from ICBC registry. Evaluating three classification models such as C4.5 DT, SVM, and ANN, describing their methodology used to extract the prediction, comparing the experimental results to estimate the model's validation and its accuracy, performing a description that can be used as criteria, and the needed procedure to make a comparison with other result.

According to the work of Ilayaraja and Meyappan [5], Apriori Algorithm was used to detect the disease existence using information in a specific period. They proposed mining approach based on rule associating and frequency generation of diseases affected by patients as well as the number of patients affected by those diseases. The dependence of this work was on training data set which gathered from electronic medical information of patients who reviewed those hospitals.

Furthermore Jain and Guatam [6] were emphasized on the efficiency of using apriori algorithm to extract the hidden patterns and generate associated rules from the data sets. They concluded in their work, that Apriori algorithm gives results that are more confident to be used for decision making by physicians and medical health institutes to discover valuable information about diseases which frequently happen.

Pushkaraj, Bhandari, Sapna, et al. [7] used Naïve Bayes model in medical data analysis which helps patients and physicians in discovering diseases at initial phases of the emergence for primitive disease relying on symptoms. This model named as predictive diagnosis system. Such systems will be more helpful in diagnosing and detecting diseases to prevent many side effects from appearing. Also they overcame the expected complexity phase of the disease as well as supported decision making for physicians and medical students to get more visions and to improve healthcare quality. The efficiency was increased and the cost was reduced.

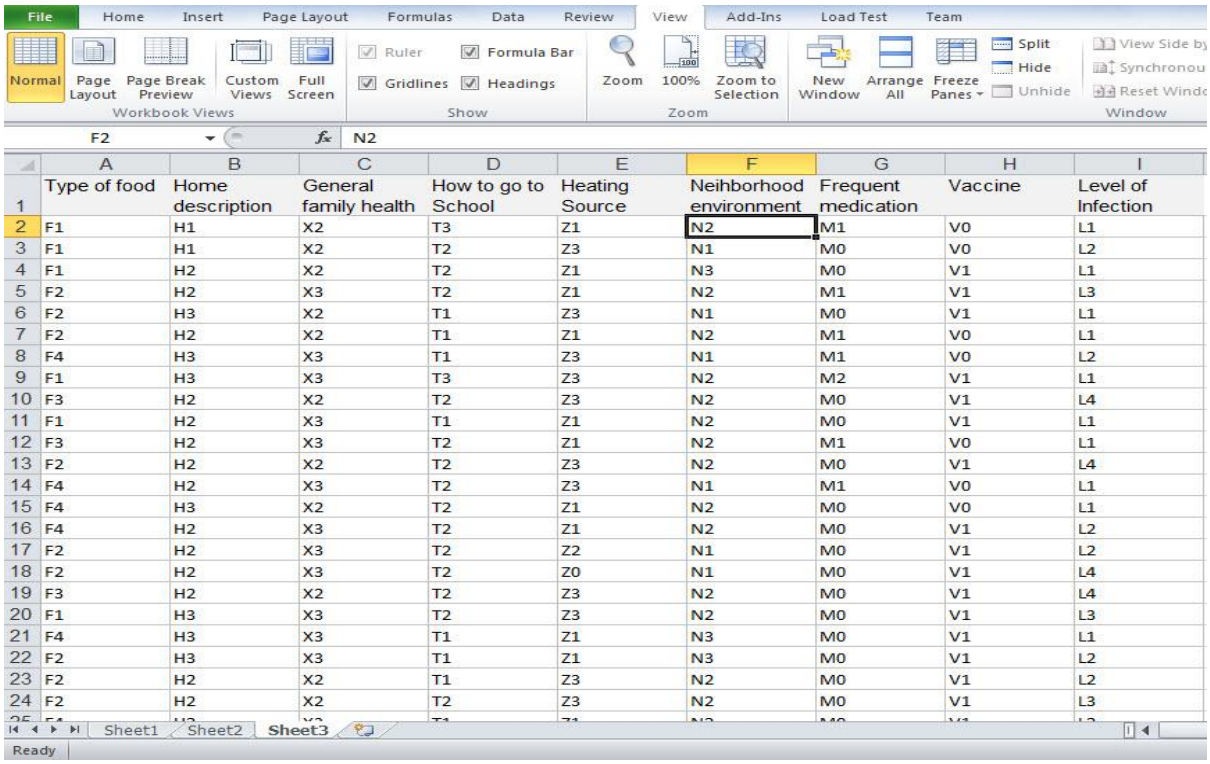
Methodology and Implementation

A consistent dataset is recommended to describe the detailed steps for extracting the hygiene information in the proposed data mining method.

An official reference book has been sent to the Iraqi ministry of education* to get there approval to distribute the 200 questioner form to the students in 2 high schools in Baghdad.

The data has been gathered and there was 149 correct forms entered in an excel sheet as in Figure-1, the data has been encoded and then transformed into WEKA form extension for further data analysis as shown in Table-1.

*Official Book Reference No. 2640 in 11\10\2016 directed from college of science\ university of Baghdad\ministry of higher education to the ministry of education\ directorate of Resafa 2.



	A	B	C	D	E	F	G	H	I
1	Type of food	Home description	General family health	How to go to School	Heating Source	Neihborhood environment	Frequent medication	Vaccine	Level of Infection
2	F1	H1	X2	T3	Z1	N2	M1	V0	L1
3	F1	H1	X2	T2	Z3	N1	M0	V0	L2
4	F1	H2	X2	T2	Z1	N3	M0	V1	L1
5	F2	H2	X3	T2	Z1	N2	M1	V1	L3
6	F2	H3	X2	T1	Z3	N1	M0	V1	L1
7	F2	H2	X2	T1	Z1	N2	M1	V0	L1
8	F4	H3	X3	T1	Z3	N1	M1	V0	L2
9	F1	H3	X3	T3	Z3	N2	M2	V1	L1
10	F3	H2	X2	T2	Z3	N2	M0	V1	L4
11	F1	H2	X3	T1	Z1	N2	M0	V1	L1
12	F3	H2	X3	T2	Z1	N2	M1	V0	L1
13	F2	H2	X2	T2	Z3	N2	M0	V1	L4
14	F4	H2	X3	T2	Z3	N1	M1	V0	L1
15	F4	H3	X2	T2	Z1	N2	M0	V0	L1
16	F4	H2	X3	T2	Z1	N2	M0	V1	L2
17	F2	H2	X3	T2	Z2	N1	M0	V1	L2
18	F2	H2	X3	T2	Z0	N1	M0	V1	L4
19	F3	H2	X2	T2	Z3	N2	M0	V1	L4
20	F1	H3	X3	T2	Z3	N2	M0	V1	L3
21	F4	H3	X3	T1	Z1	N3	M0	V1	L1
22	F2	H3	X3	T1	Z1	N3	M0	V1	L2
23	F2	H2	X2	T1	Z3	N2	M0	V1	L2
24	F2	H2	X2	T2	Z3	N2	M0	V1	L3

Figure 1- The excel sheet data before transformation

By using the General hygiene information that has been gathered from the present questionnaire, the proposed approach in this research involves the execution of three data mining algorithms consequently. Result generated from each algorithm is used as an input data for the next algorithm implementation.

The three algorithms used in the proposed approach are consequently: Apriori algorithm, association rule mining and Naïve Bayesian.

Table 1- The research attributes with their instances and symbols.

Seq.	**Attribute description(in 1 year)**	**Instance description**	**Symbol**
1	Level of Infection	Few times Average Many Too much None	L1 L2 L3 L4 L0
2	Your typical food	Starches Confetti Proteins Vegetables and Fruits	F1 F2 F3 F4
3	Home description	A small apartment Large house Large house and garden	H1 H2 H3
4	General family health	Weaker than my health Same as my health Better than my health	X1 X2 X3
5	How to go to School	Crowded bus Walk Private Car	T1 T2 T3

6	Heating source	Electric LPG Oil None	Z1 Z2 Z3 Z0
7	Neighborhood environment	Smokey Normal Healthy	N1 N2 N3
8	Frequent Medication	Pain killer and Antipyretic Antibiotic None	M1 M2 M0
9	Flu Vaccine	Yes No	V1 V0

Step1

Using the raw data collected from the schools questionnaire mentioned earlier in the dataset section, coding them and then transforms them into WEKA form performed first then the Apriori function is executed on the data to extract the frequent set and the super set as in Table- 2.

Implementing the Apriori algorithm over the health data gathered depends on a frequency threshold estimated by 9 and above as minimum frequency. We got a huge amount of data results acceding 2600 frequent item set depending on the last result. On the proposed approach 8 frequent items were extracted to represent the superset depending on two rules:

1- Selecting the super set found on the last pruning level which is the 9^{th}.

2- Selecting the most frequent item sets that override the estimated frequent threshold equal and above 30.

Table 2- The Apriori results

Apriori level	Frequent instances selected	Frequency
Level 9	F4 H2 X2 T2 Z3 N2 M0 V0 L1	9
Level 9	F3 H2 X2 T2 Z3 N2 M0 V1 L1	13
Level 5	X2 T2 Z3 N2 M0	34
Level 5	H2 X2 Z3 N2 M0	30
Level 5	X2 T2 N2 M0 L1	35
Level 5	H2 T2 Z3 N2 M0	40
Level 5	H2 X2 T2 N2 M0	40
Level 5	H2 X2 T2 Z3 M0	40

Step 2

Using the previous output in step 1 (the Apriori result) as an input in this step to catch the relations and the associations that were found in the previous super set upon estimated threshold for Support >=50% and Confidence>=70% for each pair of attributes found in the Apriori result.

Among 51 combinations for each pair of attributes found after extracting the rules, 3 rules have been selected that meet the requirements of the estimated threshold in this research as follows:

Rule 1: M0 → T2, Support = 51%, Confidence = 72%.

Rule 2: T2 → H2, Support = 50%, Confidence = 80%.

Rule 3: V1 → M0, Support = 50%, Confidence = 88%.

Step 3

The last three rules that were selected from the previous step after applying association rule mining algorithm will be used as an input data for Naïve Bayesian module that was executed in WEKA application as in Figure- 2.

Using the 4 left phases (H2=Same as my health, T2=Walking to school, M0=No medication, V1=Flu vaccination) chosen from the main remaining attributes (H=General family health, T=Transportation to school, M=Frequent medication used, V=Vaccination status) all together in one Naïve Bayesian equation. Choosing one of the attribute as a class attribute and getting all its instance results.

Repeat the above steps by changing the class attribute subsequently for the 4 attributes and finally monitor the results.

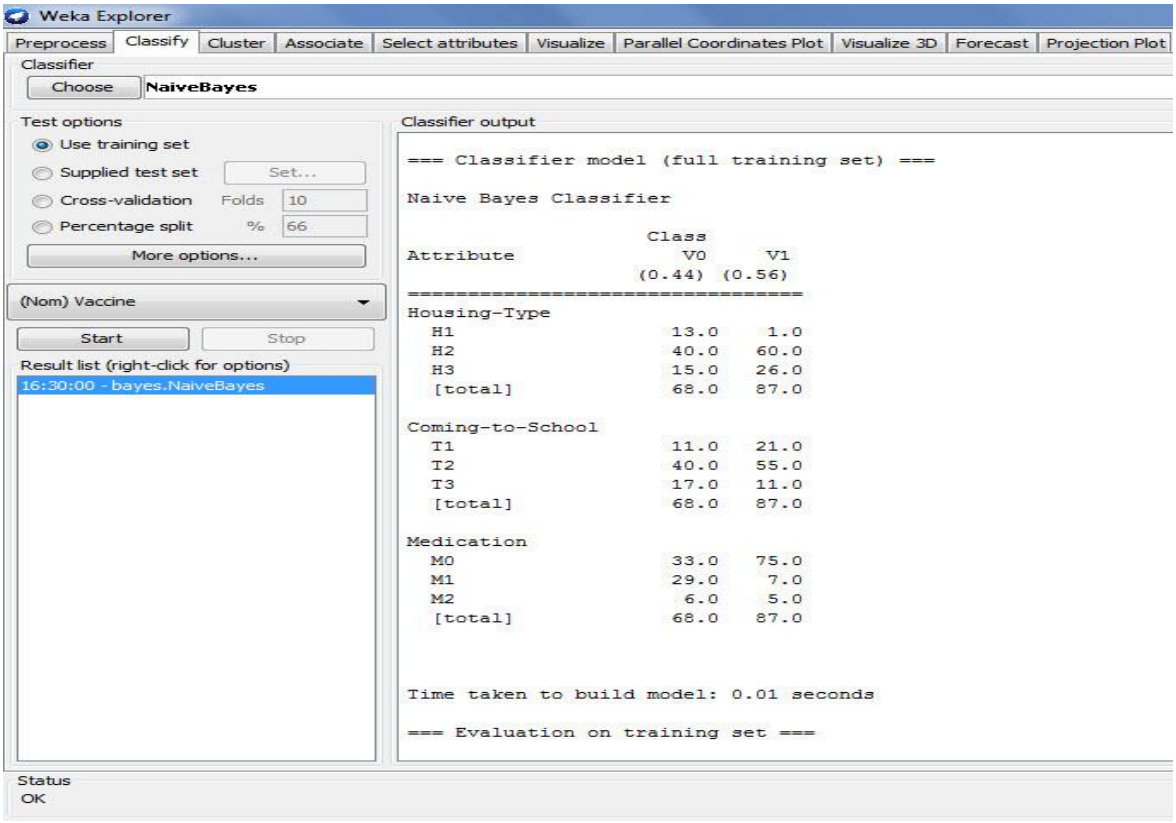


Figure 2- Executing the Naïve Bayesian in WEKA

Note the results in Figure-3, the highest value for each crosstab probability certifying the correct selection made in step 2 in association rule mining and adding multi-dimensional relation more sophisticated rule than the paired results in the association rule.

Results and Summary

Considering the results in Figure-3, notice the crosstab values at the 3rd algorithm (Naive Bayesian) with respect to the vaccination attribute as a class attribute for both cases (V0=non vaccinated, V1=Vaccinated) and the corresponding attribute set (H2=Same as my health, T2=Walking to school, M0=No medication). It is worth mentioning that the highest values in the crosstab table are the same rules that have been found in the results at the second step "the association rule". Also it is noticeable that these values were derived from the results of the first step "Apriori" of the proposed approach.

The results show some important relations that were never been in the mind of the researcher; also the whole designed questionnaire focuses on the vaccination procedure for personal hygiene. But the mining algorithms approach that has been proposed to discover new set of relations that are more likely sociology issues but rather interesting.

The results showed that there is a high relation between the students that were interested in the flu vaccine and there life style, these students mostly don't use any kind of medication (i.e. pain killers or antibiotics) those specific students also don't use transportation when they go to school, they choose to come walking.

Naive Bayes Classifier

Attribute	Class H1 (0.09)	H2 (0.65)	H3 (0.26)
Vaccine			
V0	13.0	40.0	15.0
V1	1.0	60.0	26.0
[total]	14.0	100.0	41.0
Medication			
M0	7.0	72.0	30.0
M1	7.0	23.0	7.0
M2	1.0	6.0	5.0
[total]	15.0	101.0	42.0
Coming-to-School			
T1	1.0	14.0	18.0
T2	7.0	75.0	14.0
T3	7.0	12.0	10.0
[total]	15.0	101.0	42.0

Figure 3- Naïve Bayesian crosstab results

Conclusions

The hygiene field researchers maybe specialized in their own scientific fields, but they may be unaware of the tools, algorithms and applications of data mining that will support the decision making for physicians and medical students, getting more visions and to improve healthcare quality. The following represents the concluded results that were found in the proposed work:

- The consequently steps that were implemented in the proposed work conduct a direction guide for the researchers of general hygiene field to be followed easily and logically.
- Discovering unknown relationships and connections among the attributes under consideration that even the researchers themselves didn't plan to analyse or predict any relation about them while designing the experiment.
- These steps in the proposed approach will produce better understanding to the facts and therefore will lead to better required and diverted results that could not be extracted using the primitive and statistical common tools which were mentioned earlier.

References

1. Patel,S. and Patel, H. **2016**. Survey of Data mining Techniques Used in Healthcare Domain. *International Journal of Information Sciences and Techniques (IJIST)* **6**: 53-60
2. Bajaj, P. Choudhary, K. and Chauhan,R. **2015**. Prediction of Occurrence of Heart Disease and Its Dependability on RCT Using Data Mining Techniques. *Information Systems Design and Intelligent Applications*, **340**: 851-858.
3. Etter,P.V., Huttunen,S., Vihavainen,A., Vuorinen,M. and Yangarber,R. **2010**. Assessment of utility in web mining for the domain of public Health. Proceedings of the NAACL HLT 2010 2nd Louhi Workshop on Text and Data Mining of Health Documents, Los Angeles, California, USA, 1-6 June.
4. Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M. and Razavi, A.R. **2013**. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, **4**:124-133.

5. Ilayaraja, M. and Meyyappan, T. **2013**. Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm. Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), Periyar University, Salem, India, February 21-22.
6. Jain, D. and Gautam, S. **2014**. Implementation of Apriori Algorithm in Health Care Sector: A survey. *International Journal of Computer Science and Communication Engineering,* **2**(4): 26-32.
7. Pushkaraj, R.B., Sapna, P.Y., Shyam, A. M., Devika, P.R. **2016**. A Survey on Predictive System for Medical Diagnosis with Expertise Analysis. *International Journal of Innovative Research in Computer and Communication Engineering,* **4**(2): 2026-2029.