ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

A NOVEL APPROACH FOR ANALYSIS OF BREAST CANCER AND MENTAL HEALTH USING VARIOUS DATA MINING TOOLS

S. Arul Murugan[1], M. Kannan[2]

Research Scholar, Dept of CSA, SCSVMV University[1]

Asst. Prof, Dept of CSA, SCSVMV University[2]

Abstract: Data mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data. In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available. But the large amount of data cannot be processes by human being in the limited period. So we are using various data mining tools to process these data in health care efficiently and in the timely manner. We can also have expected outcome rather than manual calculation. In this dissertation we consider two types of medical database which one shows the analysis of breast cancer using various data mining and another one for mental health as a case study. Here the breast cancer consists of 10 attributes whereas minimum data set for mental health having 455 attributes. In this dissertation we consider some of the tools such as weka, cruise, discover E to evaluate and to develop classification rule rather than using some data mining algorithms. We can have most probably 70% to 80% accurate result. A further extension of the dissertation is to store this data into PDA and storing some of the classification rule to classify the amount of data which are available to provide real time assistance to reactionaries.

I. INTRODUCTION

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year. The ability to use these data to extract useful information for quality healthcare is crucial.

Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place. It is known that "Discovery of HIV infection and Hepatitis type C were inspired by analysis of clinical courses unexpected by experts on immunology and herpetology, respectively".

Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. Imagine a doctor who has to examine 5 patient records; he or she will go through them with ease. But if the number of records increases from 5 to 50 with a time constraint, it is almost certain that the accuracy with which the doctor delivers the results will not be as high as the ones obtained when he had only five records to be analyzed.

Structured query languages (SQL) are well known software tools with very little freedom for manipulations and SQL is useful for finding information, as long as the user knows perfectly what he or she is searching for. Once the user provides the Query the processor will provide the user with the exact answer that is required for the solution. Sometimes we come across cases where the patient has symptoms of fever and sweating. SQL cannot provide us with a diagnosis or decision about whether the patient is having a headache or a cold based on the information provided.

This lead to the use of data mining in medical informatics, the database that is found in the hospitals, namely, the Hospital Information Systems (HIS) containing massive amounts of information which includes patients information, data from laboratories which keeps on growing year after year. With the help of data mining methods, useful patterns of information can be found within the data, which will be utilized for

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

further research and evaluation of reports.

The other question that arises is how to classify or group this massive amount of data. Automatic classification is done based on similarities present in the data. The automatic classification technique is only proven fruitful if the conclusion that is drawn by the automatic classifier is acceptable to the clinician or the end user.

In this dissertation we deal text data. A few of these problems like automated classification or diagnosis can be solved with the help of context based text classification. Typical approaches extract features out of the data that is submitted. These features are provided to machine learning with the help of pattern extraction techniques. These features usually include some patterns or words that can be used to extract the other words or patterns relevant to the end user, which will help to categorize the data.

However, in this dissertation we look at various data-mining tools, as all data is considered as simple data, to perform automatic classification based on the testing data set and also provide accuracy in terms of percentage with regard to the number of cases in the testing dataset, that were classified correctly.

In both case studies presented in this dissertation we know the categories or outcome with respect to the different cases, thus we will concentrate mainly on supervised learning methods in data mining. Suppose information regarding classification or outcomes of the cases was not present, the result would be the use of unsupervised learning methods.

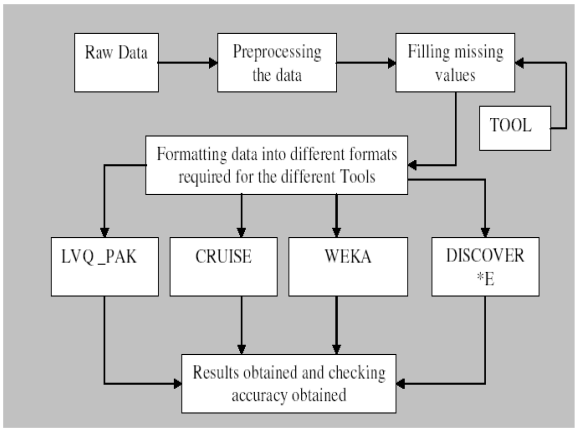
Although none of the data makes any sense to the compiler or the machine learning algorithms, text data are rather easier for classification and categorization than other types of data. Also with text data, results are more accurate and are obtained more quickly than with other types of data.

With mobile computing dominating the market it is possible to build software on mobile or hand held devices such as a PDA or a smart phone. These devices are handler than laptops and allow for easier access at all times. The drawback of today's PDAs is that they have low computing power and small storage capacity. Thus, running these algorithms on PDA is not feasible due to these factors.

Lastly, some of the data mining algorithms make use of rules, which are required for categorization. Rules are obtained based on patterns present in the training data set, which are extracted by the various data mining algorithms. This rule-based stage can be performed on a desktop. Once these rules are obtained they can be stored on a PDA. Inputs regarding the patient can be fed to the PDA and classification of the input can take place based on the rules stored in the device in real time.

II. SYSTEM ARCHITECTURE

The different components of the systems are as connected as shown in the following figure. The flow of the system starts with the collection or raw data, which is used for data mining. This data is first preprocessed by the different tools and converted into formats understood by the different tools that are used in the mining process. Missing values can be either filled in the preprocessing stage or by using a separate tool, for example as the one shown in the WEKA software, explained later. The training part of the cleaned data is first passed into the different data mining tools where similarities in the patterns are extracted. Once these similarities in the data are extracted they can be called as patterns or rules.



System Architecture

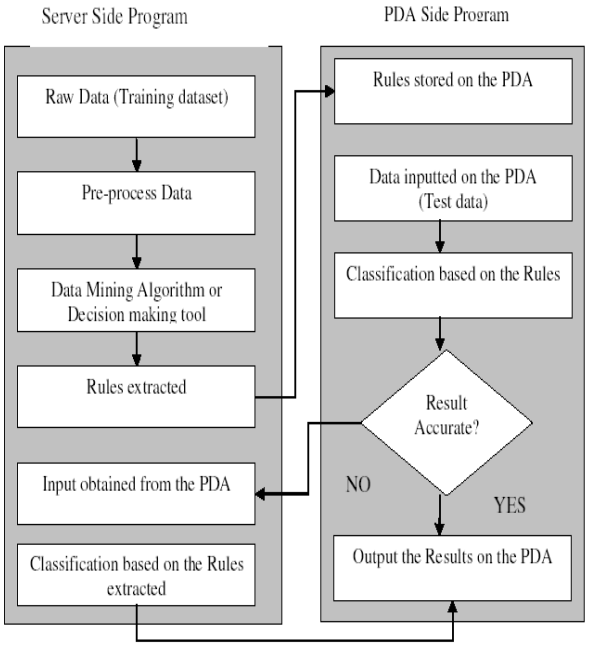
An objective of this dissertation was to develop a tool that can be developed on a handheld or a mobile computing device, such as a PDA. We can implement these tools to work well on a computer say a desktop or a laptop, but integrating the same tool on a hand held can be rather tricky. The drawback of this type of device is that they have low memory and low computational power.

Thus, instead of storing all the data and the data mining algorithms on the tool, Handheld device, we run these tools on desktop computers and save only the inference engine or the rule set on the PDA. We then input the data directly on the PDA and the rule set can be run to provide the required answer. When there is need for large computing power, with the help of an Internet service, we can send the data to the server where computation can take place and output the results from the server to the PDA.

Thus the architecture of the system with the PDA in mind is as shown below.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013



Detailed architecture of the system

III. DATA PREPROCESSING

Each algorithm requires data to be submitted in a specified format. The generation of raw data into machine understandable format is called preprocessing. Other steps that are performed during preprocessing are the transformation of the attributes in the database into a single scale and the replacement of all the missing values in the data.

Machine understandable format

Raw data can be stored in several formats, including text, Excel or other database types of files. Sometimes the raw data is not in any format.

Having data already in a format understandable by algorithms can result in better time efficiency with respect to processing of the data. In most cases the rows represent a single case and columns represent the attributes that are present within this case. In some of the free databases that are available online most of them are in comma separated value (CSV) format. That is all the attributes are separated by commas and two commas simultaneously stands for a missing data attribute. Sometimes when attributes are missing, instead of finding an empty space we may find a question mark in place of the missing attribute.

In the WEKA tool for example, the data should be stored in the Attribute-Relation File Format (.ARFF format) as the data type of the attributes must be declared. The system does not automatically classify the attribute as being real or categorical. An example of the ARFF format will be described in the next section.

The Wisconsin breast cancer database is described below to illustrate how the preprocessing is done to provide inputs to each of the machine intelligent tools that were used.

Raw data

The raw data usually has a great deal of noise. Raw data cannot be used directly for processing, with the machine-learning algorithms. They first need to be preprocessed into machine understandable format.

Machine understandable format in WEKA

Most data mining tools can use data in the CSV format for running the machine intelligent algorithms. The data that is used for WEKA should be made into the following format shown in the table below and the file should have the extension dot ARFF (.arff). The last attribute where the classification of the patient is done is made into a categorical format, that is, the classification attribute ‘diagnosis’ takes string values ‘a’ when cancer is beginning and ‘b’ when cancer is malignant. The missing values are replaced by ‘?’ mark.

```
@relation 'cancer'
@attribute 'ClumpThickness' real
@attribute 'UCellSize' real
@attribute 'UCellShape' real
@attribute 'MAdhesion' real
@attribute 'SEpithelialCellSize' real
@attribute 'BareNuclei' real
@attribute 'BlandChromatin' real
@attribute 'NormalNucleoli' real
@attribute 'Mitoses' real
@attribute 'Diagnosis' {'a','b'}
@data
6,8,8,1,3,4,3,7,1,a
4,1,1,3,2,1,3,1,1,a
8,10,10,8,7,10,9,7,1,b
```

Machine understandable format in CRUISE

Two files are required for the compilation of the database with respect to the CRUISE software. One file contains the description of the attribute and the other file consists of all the data that is present in the database. In the description file “bcancerwis.txt”, is the file where the data is located and ‘?’ is used as a code for missing values. The rest of the data consists of information about the different attributes, e.g. ‘c’ in vartype means the attributes is categorical. In these cases ‘n’ means the attribute is numerical and ‘d’ means that the attribute is dependent and so on.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

The description file appears as follows

bcancerwis.txt

Column	varnam	vartype
1	ClumpThickness	n
2	UCellSize	n
3	UCellShape	n
4	MAdhesion	n
5	SEpithelialCellSize	n
6	BareNuclei	n
7	BlandChromatin	n
8	NormalNucleoli	n
9	Mitoses	n
10	Diagnosis	d

The data file is a CSV format file

6, 8, 8, 1, 3, 4, 3, 7, 1, a
4, 1, 1, 3, 2, 1, 3, 1, 1, a
8, 10, 10, 8, 7, 10, 9, 7, 1, b

The data used as input in CRUISE looks similar to the one used in WEKA. The difference between the two is that, in WEKA the descriptive file of the attributes is present within the dataset and in the case of CRUISE there are two files which need to be inputted to the tool, one containing the description of the attributes and another containing the dataset as shown above.

Machine understandable format in Discover E

For the Discover E tool the data is provided in a similar format as the CSV file with the name of the attributes at the first line of the data set. This data set is first sent through the Importer tool which automatically converts the data into the machine understandable format for the Discover E tool. The file that is created has a dot mining (.mining) as the extension of the processed file.

Preprocessor tool that is present in Discover E software

Unlike the other tools, the data need not be stored in a particular format. The data, which is provided above, is in the CSV format with "?" representing the missing data in the database. This tool makes the data into a format suitable for this tool to provide data analysis easily.

Machine understandable format in Learning Vector Quantization

In the LVQ the data presented to the tool is not in the CSV format. The attributes are separated by space and the missing value is represented by 'x'. The number of attributes that are present to make the diagnosis should also be specified. If we look at the example of the raw data given below we see that there are 9 attributes that are required for the classification attribute mentioned in the last column. Thus the number 9 has to be mentioned in the first line of the dataset, which relates to the number of attributes that are present. Also all the attributes should be given in real numbers.

The first few lines of the data looks like this:
6.0, 8.0, 8.0, 1.0, 3.0, 4.0, 3.0, 7.0, 1.0, a
4.0, 1.0, 1.0, 3.0, 2.0, 1.0, 3.0, 1.0, 1.0, a
8.0, 10.0, 10.0, 8.0, 7.0, 10.0, 9.0, 7.0, 1.0, b

IV. CONCLUSION AND FUTURE WORK

CONCLUSION

Machine intelligence algorithms are improving as the number of data mining tools and algorithms increase. Healthcare data is a good test bed for data mining. A great deal of data in health care is still being gathered and organized using pen and paper. In this dissertation, we have used the MDS-MH as the case study that consists of 455 attributes and over 4000 cases.

The minimum dataset that was analyzed is in the area of mental health. There are a number of other tools that are based on MDS and have been made mandatory in different parts of Canada. The advantage of the MDS assessment tools is that they can be integrated with each other, resulting in a much bigger set of data. Thus soon there will be a number of other integrated tools in the MDS system for data mining.

In this dissertation, we used Zero R as the base case. Sometimes it outperformed some of the other data-mining algorithms and one reason being that Zero R implements the majority class to be the output with regard to the final output of the tool.

If we can classify the testing data set into 2 categories say X and Y, and in the test data set there are more cases present in category X than Y, then the Zero R tool will be trained to predict the category for any test case as X as the tool is trained to classify all the outcomes based on the majority class. Similarly in experiment 1 in table 9, 75.75% was the accuracy obtained for Zero R method, which means 75.75% of the test data, represents the majority class of the training set. Thus the time required for computation and classification in this method is minimal.

An Example where the Zero R could perform better is, consider a case where 99 out of 100 cases belong to the majority class of the training dataset. In this the prediction rate of the Zero R tool is 99%. But incase in the testing dataset there is only one instance of the majority class of the training dataset then the prediction of this tool will be 1%. Thus the tool is completely biased on the distribution of the training dataset.

The Naive Bayes algorithm provides very fluctuating results in the MDS-MH data set. This is an algorithm commonly used to produce classified results at a very high speed. Accurate prediction with the Naïve Bayes algorithm comes when all the independent variables are

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

statistically independent of each other. Accuracy with respect to the rule based classification can be increased by using more rules for the classification of the test data.

The decision tree experiments that were conduced were the most useful and informative experiments. One of the questions was whether the number of attributes in the database could be decreased. To answer the above question we look at the experiments conducted by WEKA on using decision tree in section 4.1.1. We find here that for the breast cancer research the total number of attributes used were four out of the ten that were available which provided an accuracy of 98.995% as mentioned in Table 1.

Also for the MDS-MH data set for Experiment 9 that is provided in Appendix B and C the number of attributes that where used for the experiment were 163 out of 455 present in the database, which provided an accuracy of 80.76% as shown in Table 12. The number 163 was obtained from the tree using a Java program.

V. FUTURE WORK

Mobile computing plays a very important role in today's information retrieval system. Some of the new handheld devices, cellular phones, PDAs, the Blackberry and others can be connected to the Internet and information can be received and sent from servers.

There are a number of different data mining algorithms that produce rules that can be stored in mobile devices and used for data classification. A possibility for future work could be to implement a local interface for the device where user can input data directly into their mobile devices, and based on the rule set, can deliver the answer back, i.e. classification is done using rules stored in the database of the PDA. This can be a handy tool for medical practitioners.

REFERENCE

1. Huang, H. et al. "Business rule extraction from legacy code", Proceedings of 20th International Conference on Computer Software and Applications, IEEE, 1996.
2. Anthony S. Fauci, et al 1997. "Harrison's Principles of Internal Medicine ed. New York": McGraw-Hill
3. Tom Mitchell 1997 "Machine Learning", McGraw Hill.
4. Lloyd-Williams,M. "Case studies in the data mining approach to health information analysis", *Knowledge Discovery and Data Mining (1998/434), IEEE Colloquium on,8 May1998.*
5. IBM Guide Business Rules Project, "Defining Business Rules – What are they are really", http://www.guide.org/pubs.htm, 1996
6. ILOG Rules white paper, http://www.ilog.com/resources/whitepapers.cfm
7. Kan, S.H. 1995 "Metrics and Models in Software Quality Engineering", Addison Wesley.
8. B. Wuthrich. "Knowledge Discovery in Databases". *Technical Report CS-95-4,* The Hong Kong University of Science & Technology, 1995. http://citeseer.nj.nec.com/89234.html
9. U Fayyad, P.Shaptro and P.Smyth. "From data mining to knowledge discovery in databases", American Association of Artificial intelligence. 1996. http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf
10. Hirdes JP, Fries BE, Morris JN, et al. "Integrated Health Information Systems Based on the RAI/MDS Series of Instrument" Healthcare Management Forum 12(4):30-40, 1999