Research Article

Development of a Data Mining Tool on Android Smartphones

Chanchai Supaartagorn*
Department of Mathematics Statistics and Computer, Faculty of Science,
Ubon Ratchathani University, Warin Chamrap, Ubon Ratchathani 34190, Thailand

Abstract

Data mining is an important part of the knowledge discovery process that can be used to analyze large datasets as well as hidden and useful information. Data mining is applied effectively not only in scientific research but also in business environments. Currently, there is a wide selection of data mining tools, such as RapidMiner, R, Weka, KNIME, etc. However, these tools work with desktop or laptop computers only. Consequently, it is inconvenient to be used for mobile devices such as tablets and smartphones. This work aims to develop a data mining tool that can be used on android smartphones. The tool can upload datasets and choose data mining techniques, namely association, clustering and classification. In developing the system it was connected to the Weka to extract a model and transform it into an understandable format. In addition, the system's performance was evaluated by two groups, 3 experts and 23 general users. The results showed the average value of the satisfaction level were 4.46 and 4.17 with standard deviations at 0.55 and 0.69 for the experts, and general users, respectively. It was found that the system performance of the tool reached an agree level. It was revealed that the developed system can be used precisely and effectively.

Keywords: Data Mining; Android; Smartphone; Weka.

1. Introduction

Data mining is the process of discovering patterns and useful information in large datasets. The term "Data Mining" can be related to the metaphor as looking for gold or diamond in a mountain. It not only can be used in scientific research, but also be applied to serve business transactions, such as market segmentation, customer churn, fraud detection, direct marketing, interactive marketing, market basket analysis and trend analysis. The process of data mining is integral steps in knowledge discovery in datasets. The process includes data preprocessing and post-processing. Datasets can be stored in a variety of formats, including flat file, csv file, or relational databases. The purpose of data preprocessing is to transform incomplete and noisy datasets into an appropriate format. After data preprocessing, various modeling techniques are selected and applied, depending on the tasks goal. The purpose of post-processing is to interpret patterns. Visualization of the results from a variety of viewpoints is also encouraged.

Currently, there are many data mining tools which are available either as open-source or commercial software. A June 2014, poll published on the influential KDnuggets portal [1], regarding the use of data mining tools in a real project. The top 5 of free data mining tools are as follows: RapidMiner (44.2%), R (38.5%), Python (19.5%), Weka (17.0%) and KNIME (15.0%). Although these tools are supported on multiple operating systems (Windows, OS X, Linux), they can be used with (desktop or

***Correspondence :** chanchai.s@ubu.ac.th

laptop) computers only. The purpose of this work is to develop a data mining tool that can runs on smartphones as a mobile application.

The computer world is moving to the Post-PC era. Smartphones can do almost everything a personal computer can do, due to their powerful on-board computing capability, capacious memory, large screens and open operating systems that encourage applications development [2]. Now, mobile devices have replaced PCs as the primary means to access the Internet. Far more Internet-capable mobile devices are now sold as opposed to PCs. Smartphones are taking over many of the functions that PCs traditionally performed [3]. According to a report by eMarketer [4], the global smartphone audience surpassed the 1 billion mark in 2012. Smartphone users increase by 22.5% to 1.75 billion users in 2014 and smartphone uses worldwide will total 2.5 billion in 2017.

Android is the leader of mobile OS market share, with 283 million units shipped and over 84% of the market share in the third quarter of 2014 [5]. In addition, there were 1.4 million Apps available in the Google Play Store in February 2015 [6]. The Android platform has also grown to become favorite among mobile developers.

This research proposed a data mining tool developed for android smartphones by employing three techniques: Association (Apriori algorithm), Clustering (K-means algorithm) and Classification (Decision tree - C4.5). The tool provides an output that is easy to use and understanding. Users can upload datasets and store outputs in their account. The tool was developed with Android Studio and PHP language, which connect with Weka for discovering patterns and useful information. Moreover, a MySQL database was used for storing datasets.

The rest of this paper is organized as follows: Section 2 describes the related works. Section 3 presents the system architecture. Section 4 shows the implementation of the data mining tool. Section 5 discusses the system evaluation. Section 6 draws the conclusion and proposes for future research.

2. Related Works

Data mining is widely used in diverse areas. There is a great deal of research in data mining systems available today and yet there are many challenges in this field. For example, an education system [7] that mined in educational environment is called Educational Data Mining. This research used K-means clustering algorithm to predict academic performance of students. The model grouped data into three clusters according to the final grades: low, average and high. This study helps teachers to reduce drop-out ratio to a significant level and to improve the performance of students. E-commerce system [8] refers to the buying and selling of goods through electronic media such as the Internet. This research used the decision tree technique to classify mobile phone ratings. The attributes present in the mobile phone datasets are seller rating, product's company, model number and product price. There are three classes of product rating: best, medium and low. This study uses analysis to aggregate and summarize the feedback from the customers which is available online as product reviews. Heart disease diagnosis [9, 10] is the process of determining which heart disease or condition explains a person's symptoms and signs. This research proposed to determine the heart diseases through various data mining techniques. The data mining could help in the identification or the prediction of high or low risk heart diseases. The most important attributes for heart disease are age, sex, chest pain, blood pressure, personnel history, previous history, cholesterol, fasting blood sugar, resting ECG, Maximum heart rate, slope, etc.

Then, we reviewed the data mining technique, which provides three techniques available on the data mining tool.

- Association is the method to discover the relationship of a particular item in a data transaction on other items in the same transaction. It is intended to predict patterns. For example, if a customer purchases a laptop PC, then customer also buys a mouse in 60 percent of the cases [11]. The Apriori algorithm calculates rules that express probabilistic relationships between items in frequent itemsets. For example, a rule derived from frequent itemsets containing A, B, and C might state that if A and B are included in a transaction, then C is likely to also be included.
- Clustering refers to the grouping of records, observations, or cases that are similar in the same categories and dissimilar in other categories [12, 13]. For example, clustering tasks in business, customer transaction data can be mined to identify market groups or consumer satisfaction. This is very important for business support and decision making. Typically, we wish to cluster data into a specified number of clusters, as in the case of the well-known K-means algorithm, which is centroid model. K-means clustering aims to partition n objects into k clusters. The algorithm determines the centroid coordinate and distance of each object to the centroid. Then, the algorithm groups the objects based on minimum distance.
- Classification is concerned with the construction of 'classifiers' that can be applied to 'unseen' data in order to target categories or classes [12]. The goal of classification is to predict the target class accurately for each case in the data. For example, a bank loan officer can use to identify loan applicants as low, medium or high credit risks. A doctor can use to diagnose categories of heart disease patients as normal or abnormal. Decision tree is the simplest model of classification. It uses a tree-like graph to model the training set. In the tree, inner nodes represent attributes and leaves nodes represent the class. C4.5 (J48 on Weka) is a well-known decision tree algorithm, which is an extension of ID3 algorithm. It induces decision trees and generates rules from datasets, which may contain categorical and/or numerical attributes. The rule can be used to predict categorical values of attributes from new records [14].

Next, we reviewed free software tools for general data mining, which is the information to decide for developing of a data mining tool. RapidMiner, R, Weka and KNIME were showed the brief details as follows:

- RapidMiner was developed by the company RapidMiner, Germany. The tool offers an integrated environment with a visually appealing and user-friendly GUI. Processes contain operators in the form of visual components. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators.
- R is an open source software platform and programming language for statisticians and data mining tasks. R offers very fast implementations of many machine learning algorithms. The main problem with R is its language, which although highly extendable, is also difficult to learn and understand thoroughly enough to become productive in data mining.
- Weka was developed by the University of Waikato, New Zealand. The tool has become very popular and has a large community for support, which is due to its user friendliness and the availability of a large number of implemented data mining algorithms.
- KNIME was developed by the Swiss company, KNIME.com AG. The tool integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing for modeling, which consist of nodes that process data, transported via connections between those nodes.

These tools have implementations for Windows, Linux and Mac OS X operating

systems. The supported data mining algorithms are shown in Table 1.

Table1. Data mining algorithms by the tools, adapted from [2]**.**

Category	Rapid Miner	R	Weka	KNIME
Decision tree (C4.5)	A(Weka)	A RWeka	+	-
Clustering (K-Means)	+	+	+	+
Association rules (Apriori)	A(Weka)	A RWeka	+	A (Weka)

The tools either implement an algorithm (+), use an external add-on (A) to support it, or do not implement it (-) at all. RapidMiner, R and KNIME use an external add-on from Weka and RWeka. Therefore, we decide to use Weka for developing this data mining tool. The application sends the input data and datasets to Weka, which it processes and shows the output in an easy to understand format.

3. System Architecture

Data mining tool are a Three-Tier system, which means that it is a client-server architecture in which the user interface (Presentation Tier), function process logic (Logic Tier) and data storage (Data Tier). Fig. 1 illustrates the architecture of data mining tool.

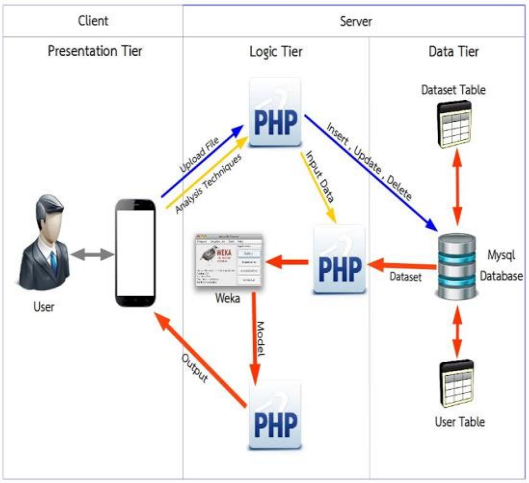

Fig.1. Architecture of data mining tool.

- Presentation tier provides the user interface on an android smartphone. The user interface and functions were developed with Android Studio. The data mining tool application was installed on the android smartphone, which can be used at remotely via the Internet. The details of this layer start from user login via the smartphone. The user sends a request to the server, namely datasets and data mining technique. The datasets come from data preprocessing process that involves transforming raw data into an understandable format. The format of datasets is Attribute-Relation File Format (ARFF) or Comma-Separated Values (CSV). The datasets will be downloaded from outside into smartphone. Then, the datasets can be uploaded via smartphone to the server. The presentation tier forwards the request to the logic tier that will perform and coordinate the application. After that, the output is reported back to the user.
- The logic tier is the coordinator between presentation tier and data tier. It also processes commands, makes logical decision and performs calculations. There are three main functions in this layer: data manipulation, input data and Weka integration. The data manipulation is the process for inserting, updating and deleting datasets. The android application can be connected to MySQL database using PHP language. The input data is the process for selecting the options of data mining technique to extract the output that fulfill specified options. Weka integration performs the connection between PHP language and Weka. PHP uses exec() command to execute the Weka processing, e.g. the command for clustering technique with weather dataset.

$cmd = "java –cp weka.jar weka.clusterers. SimpleKMeans – N 2 –t data\weather.arff"; exec($cmd, $output);

After that, PHP retrieves the output model and sends back to the android smartphone in JSON format (JavaScript Object Notation),

e.g. the output model of clustering technique in JSON format.

[
["age","50-59","40-49","40-49","50-59","50-59\n"],
["menopause","premeno","premeno","premeno","ge40","ge40\n"],
["tumor-size","30-34","20-24","30-34","25-29","30-34\n"],
["inv-nodes","0-2","0-2","0-2","0-2","3-5\n"],
["node-caps","no","no","yes","no","yes\n"],
["deg-malig","2","2","3","1","3\n"],
["breast","left","right","right","left","left\n"]
,
["breast-quad","left_low","left_up","left_up","left_low","left_low\n"],["irradiat","no","no","no","no","yes\n"],
["Class","no-recurrence-events","no-recurrence-events","recurrence-events","no-recurrence-events","recurrence-events\n"]
]

The JSON format will be transformed to display on android smartphone is shown in Fig. 2.

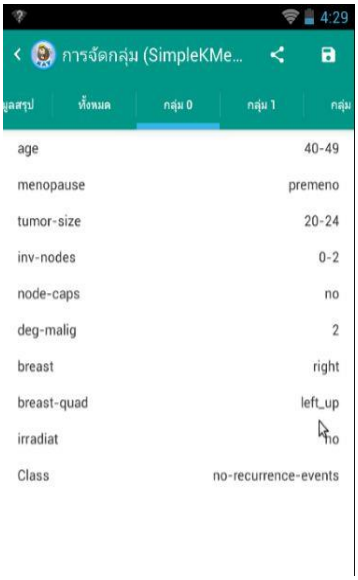

Fig.2. The output model of clustering.

- The data tier is responsible for data storage via the MySQL database. There are two tables in the database, namely the user table and dataset table. The user table stores user information for user authentication. The dataset table stores the collection of data, which was uploaded from the user.

4. System Implementation

In this section, we show the interface of an application and an example in three data mining techniques. The implementation begins when the users registers to the application and then logs on the application via the android smartphone as shown Fig. 3.

Fig.3. The logon interface.

Next, the users uploads a dataset for analyzing data. The tool determines the format of datasets as an ARFF file or CSV file. The example of file uploading is shown in Fig. 4.

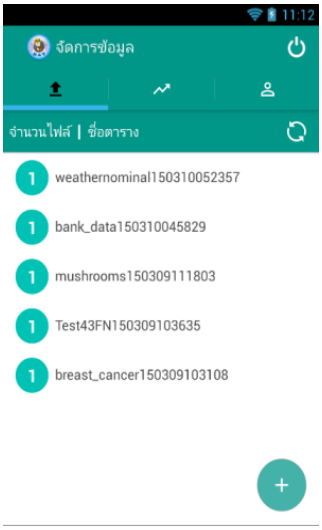

Fig.4. The example of file uploading interface.

After that, the user enters an input data options and clicks the button to show the output. An input data and an associator output are shown in Fig. 5.

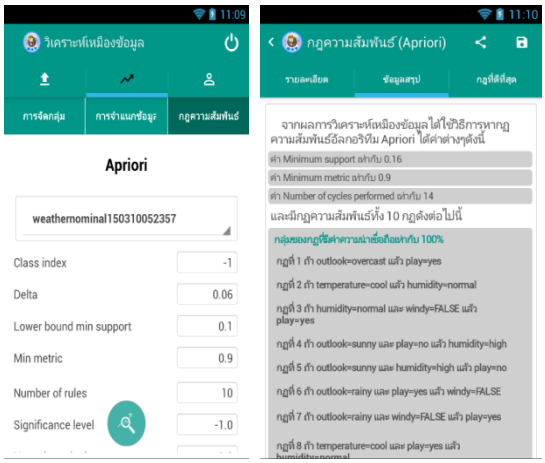

Fig.5. The example of association interface.

An input data and a clusterer output are shown in Fig. 6.

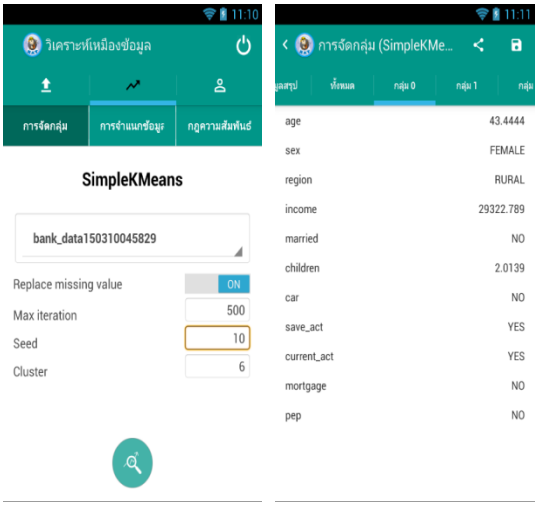

Fig.6. The example of clustering interface.

An input data and a classifier output are shown in Fig. 7.

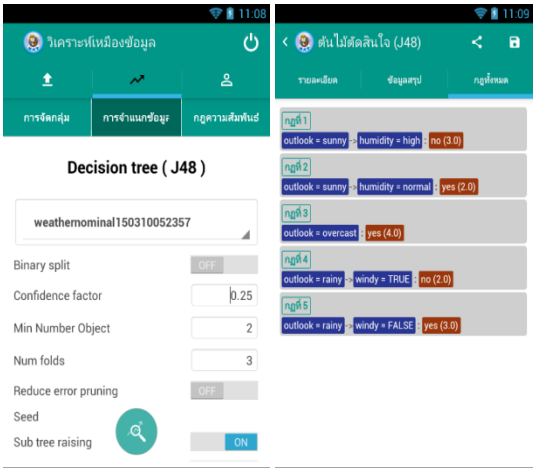

Fig.7. The example of classification interface.

In each of the above examples, mobile devices are conveniently sized to fit in small areas. Therefore, smartphone devices are not well suited for displaying text-intensive content the same way a desktop computer is. Moreover, navigation structure is also difficult to convey in the limited screen space of smartphone devices. To reduce these problems, we design user interface through TabPanel pattern where content is separated into different panes, and each pane is viewable one at a time. Compared to

traditional command line, we have grouped the information to make it easy to understand. The color scheme is used to highlight important information. The information is summarized for ease of interpretation.

5. System Evaluation

The data mining tool was evaluated by two groups. The first group consists of experts in the field of information technology, who teach Information Technology, at the faculty of Science at Ubon Ratchathani University (3 instructors). The second group is general users, who are the students studying in the Data Mining course in semester 2, academic year 2014 (23 students). The system evaluation consisted of five parts: functional requirement test, function test, usability test, performance test and security test. Five-point Likert-type questions were used in the questionnaire to assess the experts' satisfaction and the general users' satisfaction. The possible highest score was 5. The results were interpreted based on the calculations of the means and standard deviation as shown in Table 2 and Table 3.

Table2. Result of expert's satisfaction**.**

Evaluation List	$\bar{x}$	S.D.	Satisfaction Level
Functional Requirement Test	4.29	0.47	Agree
Function Test	4.73	0.48	Strongly agree
Usability Test	4.29	0.62	Agree
Performance Test	4.56	0.58	Strongly agree
Security Test	4.42	0.58	Agree
Result Summary	4.46	0.55	Agree

Table3. Result of general user's satisfaction.

Evaluation List	$\bar{x}$	S.D.	Satisfaction Level
Functional Requirement Test	4.24	0.61	Agree
Function Test	4.30	0.57	Agree
Usability Test	3.96	0.91	Agree
Performance Test	4.10	0.61	Agree
Security Test	4.26	0.73	Agree
Result Summary	4.17	0.69	Agree

From Table 2 and Table 3, the evaluation results indicated that user's satisfaction towards the system has met the needs of all users.

6. Conclusion and Future Research

In this paper we presented a data mining tool developed for use on android smartphone in order to help users discover patterns and analyze information in large datasets. The tool is convenience for mobile use, since users can use it remotely via their mobile device. The system architecture was designed based on a three-tier system as follows: The presentation tier is the user interface (UI), which displays data to the user and accepts input from the user. The logic tier handles data validation, calculation and task-specific behavior. The data tier communicates with the database, where datasets are stored and retrieved. In addition, the tool was evaluated in five parts: functional requirement test, function test, usability test, performance test and security test. The result shows that the system performance of the tool reaches an agree level. It indicates that the developed system can be used correctly and effectively. Therefore, this data mining tool can provide a solution to business and research questions that traditionally were times consuming to resolve. It helps organizations get the necessary large information needed to handle different processes as quickly as possible.

In future research, we will add more techniques to the tool, such as Classification rules, Bayesian networks, Function based

learning, etc. Moreover, we will improve the design of the tool in order to make it more user-friendly.

7. References

[1] KDnuggets, What Analytics, Data Mining, Data Science Software/Tools You Used in the Past 12 Months for a Real Project Poll, Available Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html, February 25, 2015.

[2] A. Jovic, K. Brkie and N. Bogunovic, An Overview of Free Software Tools for General Data Mining, International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, 2014, pp. 1112–1117.

[3] Thomas M.Chen, 30th Anniversary of the PC and the Post-PC Era, IEEE Network. Editor's Note, September/October. 2011, pp. 2–3.

[4] eMakketer, What Analytics, Smartphone Users Worldwide Will Total 1.75 Billion in 2014, Available Source: http://www.emarketer.com/Article/Smartphone-Users-Worldwide-Will-Total-175-Billion-2014/1010536, March 4, 2015.

[5] IDC, Smartphone OS Market Share, Q3 2014, Available Source: http://www.idc.com/prodserv/smartphone-os-market-share.jsp, March 4, 2015.

[6] The Statistics Portal, Number of Available Applications in the Google Play Store from December 2009 to February 2015, Available Source: http://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/, March 5, 2015.

[7] Bhise R.B., Thorat S.S. and Superkar A.K., Important of Data Mining in Higher Education System, IOSR Journal of Humanities and Social Science, Vol. 6, Issue 6, pp. 18-21, 2013.

[8] Aditi Todi, Anahita Agrawal, Ankit Taparia, Nikhlesh Lakhmani and Rajashree Shettar, Classification of E-Commerce Data Using Data Mining, International Journal of Engineering Science & Advanced Technology, Vol. 2, Issue 3, pp. 550-554, 2012.

[9] Aqueel Ahmed and Shaikh Abdul Hannan, Data Mining Techniques to Find Out Heart Diseases: An Overview, International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue 4, pp. 18-23, 2012.

[10] Vikas Chaurasia, et al, Early Prediction of Heart Diseases Using Data Mining Techniques, Caribbean Journal of Science and Technology, Vol. 1, pp. 208-217, 2013.

[11] Davis Olson, Yong Shi, Introduction to Business Data Mining, McGraw-Hill, 2007.

[12] Frans Coenen, Data Mining: Past, Present and Future, The Knowledge Engineering Review, Vol.26:1, pp. 25-29, 2011.

[13] Daniel T.Larose, Discovering Knowledge in Data An Introduction to Data Mining, Wiley-Interscience, 2005.

[14] Suvajit Das, Shasshi Dahiya and Anshu Bharadwaj, An Online Software for Decision Tree Classification and Visualization using C4.5 Algorithm (ODTC), International Conference on Computing for Sustainable Global Development, New Delhi, India, 2014, pp. 962-965.