EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Enhancing the performance of association rule models by filtering instances in colorectal cancer patients

Jaree Thongkam*[1)] and Vatinee Sukmak[2)]

[1)]Applied Informatics Group, Faculty of Informatics, Mahasarakham University,
Maha Sarakham 44150, Thailand
[2)]Department of Mental Health and Psychiatric Nursing, Faculty of Nursing, Mahasarakham University,
Maha Sarakham 44150, Thailand

Received June 2016
Accepted September 2016

Abstract

Colorectal cancer data available from the SEER program is analyzed with the aim of using filtering techniques to improve the performance of association rule models. In this paper, it is proposed to improve the quality of the dataset by removing its outliers using the Hidden Naïve Bayes (HNB), *Naïve Bayes Tree* (NBTree*) and Reduced Error Pruning Decision Tree (REPTree)* algorithms. The Apriori and HotSpot algorithms are applied to mine the association rules between the 13 selected attributes and average survivals. Experimental results show that the HNB algorithm can improve the accuracy of the Apriori algorithm's performance by up to 100% and support threshold up to 45%. It can also improve the accuracy of the HotSpot algorithm's performance up to 93.38% and support threshold up to 80%. Therefore, the HotSpot rules with minimum support of 80% are selected for explanation. The HotSpot algorithm shows that colorectal cancer patients, who died from colon cancer and were not receiving radiation therapy, were associated with survival of less than 22 months. Our study shows that filtering techniques in the preprocessing stage are a useful approach in enhancing the quality of the data set. This finding could help researchers build models for better prediction and performance analysis. Although it is heuristic, such analysis can be very useful to identify the factors affecting survival. It can also aid medical practitioners in helping patients to understand risks involved in a particular treatment procedure.

Keywords: Colorectal cancer survival, Apriori, HotSpot, Pre-processing

1. Introduction

Colorectal cancer (CRC) is cancer that develops in either the colon or the rectum [1]. It is a major cause of morbidity and mortality throughout the world. CRC accounts for over 9% of all cancers worldwide and affects about 5% of the U.S. population, with up to 150,000 new cases per year [2-3]. According to the American cancer society, an estimated 136,830 Americans were diagnosed with colorectal cancer, including 71,830 males and 65,000 females in 2014 [3]. In contrast, the incidence of CRC in Thailand is 13.67% of all cancers for men and 7.40% for women [4-5], with over 8,000 new cases per year [6].

Additionally, CRC is the third leading cause in mortality when males and females are considered separately, but is the second leading cause when the genders are combined. It is expected that it will have caused approximately 49,700 deaths in 2015 [7-8] and about 49,190 deaths in 2016 [8]. As can be seen, the disease affects slightly more males than females, and the risk increases with age. While the exact causes of CRC are unknown, certain factors can increase risk of the disease. These factors include lifestyle, advanced age, and genetics. Other risk factors include race, being male, high intakes of fat and red meat, alcohol, and obesity, as well as smoking, and physical inactivity [7, 9].

Although firm scientific evidence for the prevention of CRC is available, researchers continue to look for the causes of CRC as well as ways to prevent and cure the disease. In present medical studies, identifying risk factors for CRC and prediction models are normally based on multivariate statistical analysis. Big data in the healthcare system, however, contains hidden knowledge, which is impossible to discover by using conventional approaches. Data mining, therefore, is more appropriate for medical studies.

Data mining is the process of analyzing data from a massive primary data set into useful information. It is also called knowledge discovery in database (KDD). In data mining, association rule mining is a popular technique. It is the most effective data mining technique for discovering the hidden, interesting relations between variables in large databases [10]. With the help of association, we can find combinations of events that occur at the same time. Association rule mining was first introduced by Agrawal, Imielinski and Swami [11]. The process has *two key steps*:

*Corresponding author. Tel.: +66 9422 48924
Email address: jaree.thongkam@gmail.com
doi: 10.14456/easr.2017.11

finding *frequent patterns* (i.e., frequently occurring sets of items) from data and forming *association rules* [12]. A rule is measured by support and confidence to identify the most important relationships. The support value indicates how frequently the items, included in the association rule, appear in the database. The confidence value indicates how often the if/then statements have been found to be true, or, the accuracy of the association rule in the database.

Several researchers have been employing association rule algorithms such as the Apriori and HotSpot algorithms. For instance, Vinnakota and Lam [13] utilized Apriori association rule mining to investigate associations between socioeconomic characteristics and cancer mortality. They reported that health service areas with high rates of low educational attainment, high unemployment, and low income were found to relate to higher rates of cancer mortality. Agrawal and Choudhary [14] performed HotSpot association rule mining to identify factors affecting survival time. They used 13 predictor attributes to calculate lung cancer outcomes. They found that HotSpot is an effective algorithm to build association rules.

However, without improving the quality of data in pre-processing, these algorithms produce little support and confidence of the rules. Many research studies have utilized outlier filtering methods in pre-processing to improve the quality of data and performance of the prediction models. But few researchers have used filtering methods to remove outliers for improving the association rule performance. Currently, Hidden Naïve Bayes (HNB) algorithm is good at detecting network intrusion. Naïve Bayes Tree (NBTree) and Reduced Error Pruning Decision Tree (REPTree) algorithms work well in filtering outliers. Therefore, in this work, HNB, NBTree *and* REPTree algorithms are employed to eliminate the outliers in a colorectal cancer survivability data set. Additionally, we applied the *Apriori and HotSpot techniques for extracting interesting association rules and correlation relationship with the survivability of* colorectal cancer patients, using data from the Surveillance, Epidemiology, and End Results (SEER) database.

The paper is organized as follows. Section 2 deals with the methodologies, data pre-processing, the basic concepts of association rule mining, and performance evaluation applied in this paper. Section 3 shows our experimental results. Lastly, in Section 4, the discussions and further research are outlined.

2. Methodology

This section provides description of the data set, data pre-processing, association rules mining and performance evaluation.

2.1 Description of data set

The Surveillance, Epidemiology, and End Results (SEER) database for colorectal cancers (1992 to 2011) was used. SEER is an authoritative source of information on cancer incidence in the United States. SEER provides cancer incidence and survival data across several geographic regions, covering approximately 28% of the U.S population. The colorectal cancer dataset 2000-2011 was used from COLON.txt, which is provided by the SEER database. Initially, in this study, the inclusion criteria were patients with stage II, III and IV primary colon adenocarcinoma within a SEER region during the years 2000 to 2011. Exclusion criteria included cases with stages 0 and I disease, missing information and unknown values in all attributes. Also, any instances with cause of death not related to colorectal cancer were removed. The final raw data contained 39,299 instances and 13 selected attributes which were significantly related to colorectal cancer survivability in the literature [15-17], as shown in Table 1. However, we divided the sample into four subgroups by age according to K-means cluster analysis (average age at diagnosis was 68.29 years) including age group 1: 30-56 years, age group 2: 57-69 years, age group 3: 70-80 years and age group 4: 81-99 years. The average *survival* after diagnosis was *22 months*. Then the data are divided into two groups according to the average survival time. The cases of survival 22 months or more consist of 14,144 (38.11%) instances and those surviving less than 22 months, 22,971 (61.89%).

Table 1 Attributes and number of values for survival association rules of colorectal cancer

Categorical attributes	No. of values
Marital status	6
Sex	2
Race	6
Age	4
Primary site	10
Surgery primary site	3
Lymph node involvement	8
Surgery Radiation Sequence	6
Radiation	8
Scope of Regional lymph nodes Surgery	2
Grade	4
Stage	3
Cause of Death	3
Class	2

2.2 Data pre-processing

Data pre-processing refers to the tasks needed to convert the raw data into input data and is an important step in data mining. It is common that outliers exist in real world datasets. An outlier is an observation that deviates so much from other observations that it generates inappropriate changes in the overall view of the system behavior. Outlier values might arise from fraudulent behavior, human error, instrument error or simply through natural deviations in populations [18]. Several techniques can be used for outlier filtering to handle them in data sets including statistical, proximity-based, clustering-based as well as classification-based techniques [19]. Classification based outlier detection techniques assume that a classifier can be learnt from a given feature space that can differentiate between normal and outlier classes [20]. These techniques can be categorized into two groups: multi-class and one-class. Several research studies have employed classification techniques for identifying and eliminating potential outliers of mislabeled instances [21-22]. In this paper, the authors used three multi-class techniques to improve the quality of the data, including Hidden Naïve Bayes (HNB), *Naïve Bayes Tree* (NBTree*) and* Reduced Error Pruning Decision Tree (REPTree). These techniques operate under the general assumption that the training instances contain labeled instances belonging to multiple normal classes [23]. In this study, we applied the selected classifiers into 13 independent variables as well as classes. If an instance was not classified as normal by any of the classifiers, then this instance was likely an outlier. Additionally, confidence and support of association rules were utilized as criteria to compare the effectiveness of these three classification techniques.

Table 2 Number and ratio of instances before and after pre-processing

Pre-processing techniques	Total data Size	Ratio of reduction	Not survived		Survived	
			Number of instances	Ratio of instances%	Number of instances	Ratio of instances%
Raw	37,115	-	22,971	61.89%	14,144	38.11%
HNB	24,215	38.38%	19,445	80.30%	4,770	19.70%
NBTree	25,381	35.42%	19,004	74.87%	6,377	25.13%
REPTree	25,930	34.02%	19,071	73.55%	6,859	26.45%

2.2.1 Hidden Naïve Bayes (HNB)

The HNB [24] classifier is an extended form of the Naïve Bayes algorithm. The HNB model is based on the formation of another layer that signifies a hidden parent of each attribute. The hidden parent combines the influences from all other attributes. In outlier detection, HNB identified the outlier from the probability of each class in each level of hidden layer.

Paris, Affendey and Mustapha [25] reported that HNB performed surprisingly well on most classes except for the majority class, while decision trees have high accuracy on this class. Moreover, HNB is superior to *Naïve Bayes and Naïve Bayes tree*.

2.2.2 Naïve Bayes Tree (NBTree)

The NBTree, introduced by Kohavi [26], is a Naïve Bayes/decision tree hybrid. It includes Naïve Bayes classifiers as the leaves to create the decision tree, while the leaves contain *Naïve Bayesian classifiers.* An NBTree also uses a score function for assessing splitting attributes to make a tree grow and assign a class label to the NBTree. Several research studies evaluated the accuracy of NBTree. For example, Mohmood and Hussein [27] compared the performance of NBTree, Naïve Bayes and Decision Tree. The results showed that NBTree significantly outperforms Naïve Bayes.

2.2.3 Reduced Error Pruning decision Tree (REPTree)

The REPTree, a fast decision tree learner, has been proposed by Quinlan [28]. It uses a decision or regression tree and creates multiple trees in different iterations using information gained or variance for selecting the best attribute. Then, it applies a greedy algorithm and reduced-error pruning (with back-fitting) algorithm. In the REPTree algorithm, numerical values are sorted only in one round for each numerical attribute. The sorting process is for determining split points of numeric attributes. The tree is built in a greedy fashion, with the best attribute chosen at each point according to information gain.

Table 2 displays the number and ratio of instances used in this study. The raw final data contained 37,115 instances, divided into 22,971 instances for the 'Not Survived' class and 14,144 instances for the 'Survived' class. Pre-processed data using HNB consists of 19,445 instances of the 'Not Survived' class and 4,770 instances of the 'Survived' class. The pre-processed data using NBTree consisted of 19,004 instances of the 'Not Survived' class and 6,377 instances of the 'Survived' class which represented a decrease. Also, the pre-processed data with the REPTree algorithm consisted of 19,071 instances of the 'Not Survived' class and 6,859 instances of the 'Survived' class. All three filtering techniques remove outliers, which consisted of more than 30% of the data. Usually, outlier detection techniques remove only a small fraction of outliers that may distort the analysis. However, if the number of outliers in the data is large from either a data collection or data anlysis point of view, data cleaning techniques for removing large amounts of outliers are needed. Therefore, we should select only outlier detection techniques that assign each object an outlier score that characterizes the degree to which it is an outlier [29]. For example, Gamberger, Lavrac and Groselj [30] used a classification filter that removes outliers up to 52.6% of the coronary artery disease database.

2.3 Association rule mining

Association rule mining is a popular and well-known approach for extracting interesting relationships between variables in large databases. An association rule has the form of an itemset such that X → Y, where X and Y are items in the database. The X statement of an association rule is known as the antecedent, which is found in the data. The Y statement is known as the consequence of the rule and is found in combination with the antecedent. Thus, if X occurs, then Y will probably occur. The association rule is composed of two measures: support and confidence. The association rule process is comprised two steps:

1. Finding all the frequent itemsets whose number of occurrences exceeds a predefined minimum support threshold.
2. Producing rules from these frequent itemsets (with constraints of minimum confidence thresholds).

Several algorithms to extract association rules have been reported in the literature. The algorithms for association rules used in this study are dicussed below.

2.3.1 Apriori

The most popular and classical algorithm used for association rule mining is the Apriori algorithm, proposed by Agrawal, Imielinski and Swami [11]. The basic concept of the Apriori algorithm is that it uses a "bottom up" approach that extends one item at a time into frequent subsets called *candidate generation*. Then, the itemsets become groups of candidates that are tested against the data. The process is ended when no further itemset is found.

The strength of this algorithm is that it is easy to execute. It also finds all the itemsets that reach the minimum support criteria, and it can do this significantly faster than the Naïve method. However, the main drawback of the Apriori algorithm is that it needs several iterations. Several research studies have employed the Apriori algorithm. For example, Sharma and Om [31] demonstrated that the Apriori algorithm can extract association rules from oral cancer survivability database for prevention and early detection of oral cancer. Ramezankhani et al. [32] successfully used the Apriori algorithm for extracting risk patterns from a database related to Type 2 diabetes. The results showed that association rule mining is a useful method for determining which combinations of predictors occur together more often in people who will have diabetes.

2.3.2 HotSpot

The HotSpot algorithm is an association rule mining algorithm which learns a set of rules displayed as a tree structure. It maximizes or minimizes a target attribute or value of interest. The targets can be categorical or numerical. The association rule comprises of a left-hand side or antecedent and right-hand side or consequence. The antecedent identifies the segment characteristics of patients. The consequence is fixed to the target attribute, e.g., the average survival time of the patients. The HotSpot algorithm uses a greedy approach to construct a tree of rules in a depth-first search. The parameters of the HotSpot algorithm include a maximum branching factor, minimum improvement in target value and minimum segment size. The maximum branching factor refers to the amount of children of a branch, which controls the number of iterations for searching the data. The minimum improvement in target value refers to the procedure of adding a new branch based on a number setting parameter. The minimum segment size refers to the size of the segment before adding a new branch [14].

The HotSpot algorithm can generate both association rules and rule-basis for a classification problem. It is also a simple and effective algorithm for building association rules from a tree structure. A few researchers have employed the HotSpot algorithm. For example, Agrawal and Choudhary [14] used the HotSpot algorithm to generate lung cancer rules. The results showed that the rules can match the existing biomedical knowledge and provide better understanding of the risks involved lung cancer survival. Furthermore, Arikan and Gurgen [33] used the HotSpot algorithm to discover a number of rules for diagnosing cardiac problems from an ECG signal.

2.4 Performance evaluation

There are two classical measures for association rules that reflect the usefulness and certainty of a rule are support(s) and confidence(c).

Support of an association rule measures the frequency of association. It is an important measure because very low support of a rule may occur only by chance. A rule with low support is also likely to be uninteresting. Support is defined as the percentage of records that contains both X and Y among the total number of records in the dataset. Support can be expressed as:

$$Support(X \rightarrow Y) = \frac{(X \cup Y)}{N} \quad (1)$$

Confidence measures the strength or accuracy of the rule. It is defined as the percentage of data records that contain a union of X and Y to the total number of records that include X. For a given rule $X \rightarrow Y$, the higher confidence, the more likely it is for Y to be present in records that contain X. Confidence also designates an estimate of the conditional probability of Y given X. Confidence can be computed as:

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

3. Experimental results

We use the implementation of the Apriori and HotSpot algorithms provided in the WEKA Explorer Version 3.7.13 [34]. This is because the WEKA program provides a well-defined framework for experimenters to build and evaluate the models. This tool is open source and used for its many data mining and machine learning algorithms. The aim of the study is to obtain an associative data model that allows study of the influence of the input variables related to colorectal cancer survivability. In this study, we focus on results where the minimum support range is from 30% to 80% and the minimum confidence is in the sequence of support values for the three filtering techniques, as well as in the raw data. Moreover, a number of rules are presented. Also, the best technique is employed to present the association rules.

3.1 Performance of association rule models

After filtering out the outliers using the proposed methods, the reduced dataset was used for building the rule models. In this section, the confidence and support of the Apriori and HotSpot association rule models are compared with either the original or the filtered training set. Their confidence and support results are shown in Figure 1.

Figure 1 characterizes three filtering techniques as well as the raw data and compares the confidence and support factors generated from the Apriori (a) and HotSpot (b) rule models. Based on the results shown in Figure 1, HNB performed best with both the Apriori and HotSpot algorithms. It also shows that the minimum confidence level reached 100% in HNB for the Apriori and the remaining rules still had high values for this factor in the HotSpot algorithm, except for the raw dataset. Besides, the Apriori algorithm developed no rules at any support level we choose in raw data. The results about the confidence are more interesting. The Apriori algorithm produced a significant margin for minimum support values that was less than 45%, whereas for a support value of 55% or greater, the confidence values slightly decreased for the HotSpot with all filtering methods, including those using the raw data.

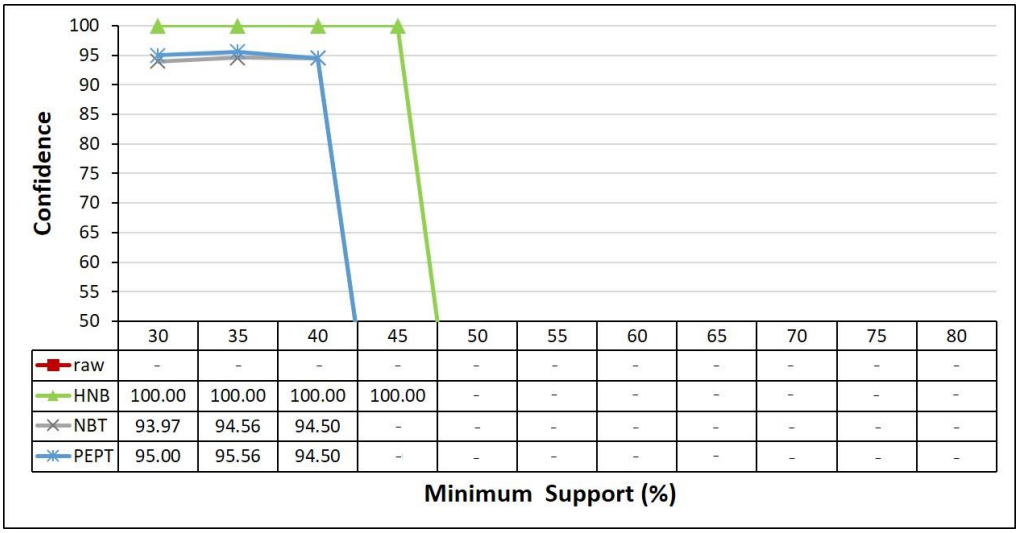
3.2 The number of rules

In this section, the number of rules generated using the Apriori and HotSpot algorithms is presented in order to verify the capability of the filtering techniques, including HNB, NBTree and REPTree. The number of rules is indicated on the y-axis while the percentages of support are on the x-axis. The results of varying the number of rules and support factors are displayed in Figure 2.

Figure 2 shows how the number of association rules increase as the minimum support is reduced using three filtering techniques. It is clear that decreasing the minimum support increases the number of rules for both algorithms. We can observe that with high support values, over 45%, Apriori cannot find any rule while HotSpot still can. Again, Apriori could not find any rules with constraints on the support in raw data. Also, it can be seen that with support values of less than 50%, the number of HotSpot rules for raw data is slightly smaller than using filters, whereas with support values over 50%, the number of rules is the same.

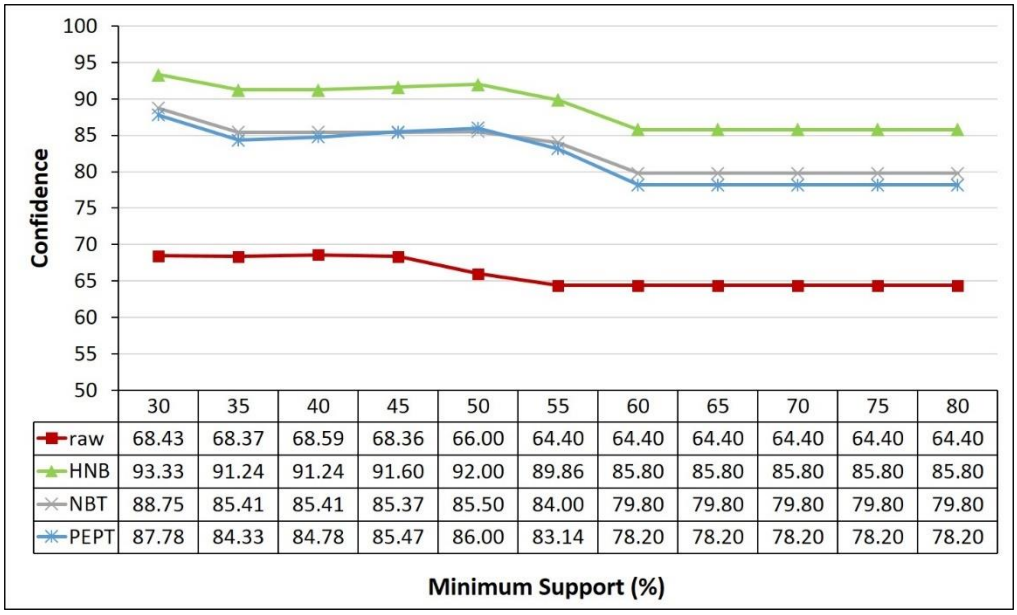
3.3 Association rules

An association rule refers to an if-then condition which is easy to interpret and understand. Our experimental results showed that the HotSpot algorithm provides high confidence and a higher support factor than the Apriori algorithm. Therefore, the HotSpot rules were chosen for this explanation. On the basis of support and confidence, the top five best rules generated by the HotSpot Algorithm using HNB (support: 80%) with significant factors are illustrated in Figure 3.



	30	35	40	45	50	55	60	65	70	75	80
raw	-	-	-	-	-	-	-	-	-	-	-
HNB	100.00	100.00	100.00	100.00	-	-	-	-	-	-	-
NBT	93.97	94.56	94.50	-	-	-	-	-	-	-	-
PEPT	95.00	95.56	94.50	-	-	-	-	-	-	-	-

(a) Performance of the Apriori Algorithm



	30	35	40	45	50	55	60	65	70	75	80
raw	68.43	68.37	68.59	68.36	66.00	64.40	64.40	64.40	64.40	64.40	64.40
HNB	93.33	91.24	91.24	91.60	92.00	89.86	85.80	85.80	85.80	85.80	85.80
NBT	88.75	85.41	85.41	85.37	85.50	84.00	79.80	79.80	79.80	79.80	79.80
PEPT	87.78	84.33	84.78	85.47	86.00	83.14	78.20	78.20	78.20	78.20	78.20

(b) Performance of the HotSpot Algorithm

Figure 1 Confidence of association models

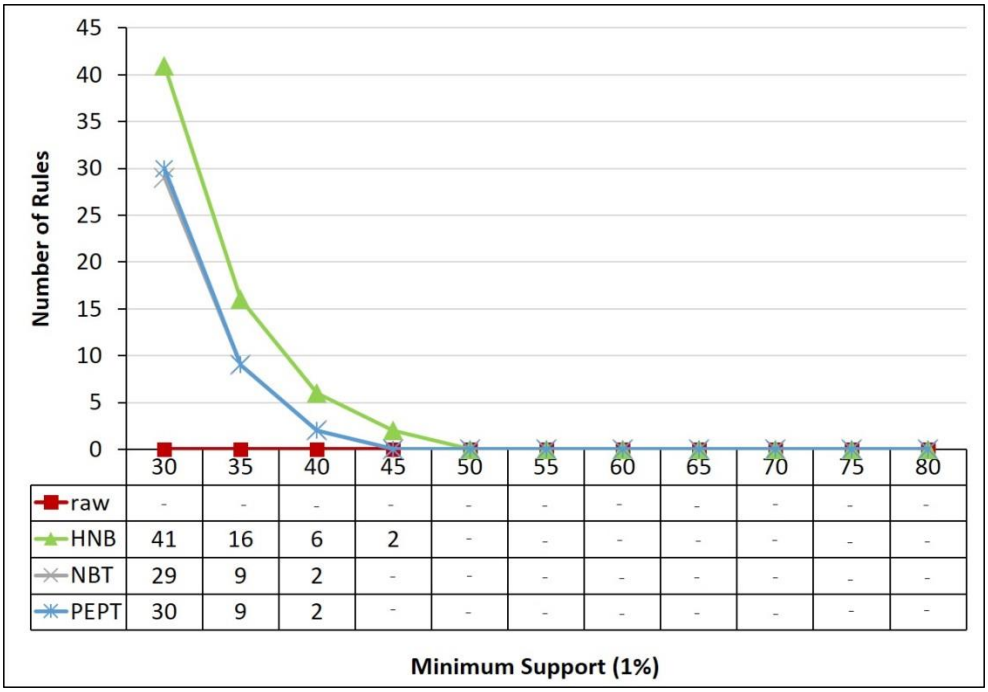
Rule Explanation: Rule 1 suggests that CRC patients (24,215 cases) who died from colon cancer excluding rectal cancer without radiation and/or cancer-directed surgery (19,252 cases: 79.50%) have the given likelihood to survive less than 22 months (16,722 cases: 69.05%) with a confidence of 87%.

Rule Explanation: Rule 2 implies that CRC patients (24,215 cases) who died from colon cancer excluding rectal cancer without radiation (18,706 cases: 77.24%) were less likely to survive less than 22 months (16,213 cases 66.95%) with a confidence of 87%.

Rule Explanation: Rule 3 implies that CRC patients (24,215 cases) who have not undergone radiation and/or cancer-directed surgery (21,768 cases: 89.89%) were the most likely to survive less than 22 months (18,818 cases: 77.71%) with confidence of 86%.

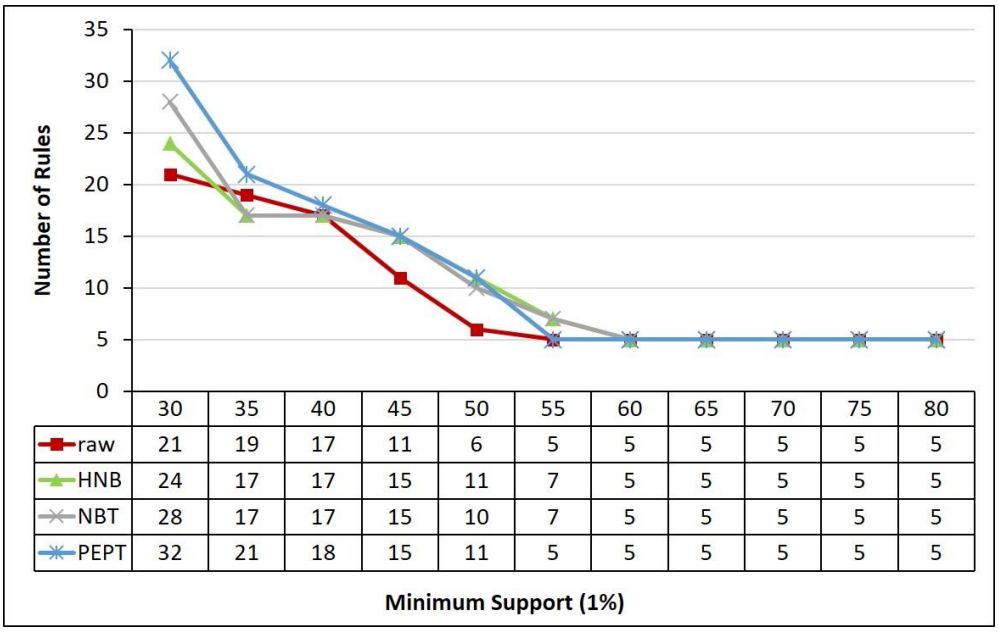
Rule Explanation: Rule 4 suggests that CRC patients (24,215 cases) who have not undergone radiation (20,596 cases: 85.05%) were likely to survive less than 22 months (17,700: 73.09%) with a confidence of 86%.

Rule Explanation: Rule 5 suggests that CRC patients (24,215 cases) who die from colon cancer excluding rectal cancer (20,566 cases: 84.93%) were likely to survive less than 22 months (17,143 cases: 70.79%) with a confidence of 83%.



	30	35	40	45	50	55	60	65	70	75	80
raw	-	-	-	-	-	-	-	-	-	-	-
HNB	41	16	6	2	-	-	-	-	-	-	-
NBT	29	9	2	-	-	-	-	-	-	-	-
PEPT	30	9	2	-	-	-	-	-	-	-	-

(a) Number of the Apriori Rules



	30	35	40	45	50	55	60	65	70	75	80
raw	21	19	17	11	6	5	5	5	5	5	5
HNB	24	17	17	15	11	7	5	5	5	5	5
NBT	28	17	17	15	10	7	5	5	5	5	5
PEPT	32	21	18	15	11	5	5	5	5	5	5

(b) Number of the HotSpot Rules

Figure 2 Number of rules comparison

Rule 1	[causeofDeath=21040, surRadseq=0]: 19252 ==> [Class=0]: 16722 <conf:0.87)>
Rule 2	[causeofDeath=21040, radiation=0]: 18706 ==> [Class=0]: 16213 <conf:0.87)>
Rule 3	[surRadseq=0]: 21768 ==> [Class=0]: 18818 <conf:0.86)>
Rule 4	[radiation=0]: 20596 ==> [Class=0]: 17700 <conf:0.86)>
Rule 5	[causeofDeath=21040]: 20566 ==> [Class=0]: 17143 <conf:0.83)>

Figure 3 Rules extracted for colorectal cancer survivability using the HotSpot algorithm

4. Discussions and further research

Outliers in a large data set can cause confusion, less accurate models and ultimately poorer results. In this paper we used three filtering techniques (HNB, NBTree and REPTree) to improve the quality of data sets for performance improvement of a rule extraction experiment using the Apriori and HotSpot algorithms for colorectal cancer survivability.

Firstly, experiments have shown that HNB performed better than NBTree and REPTree for filtering outliers to improve the quality of the data set. Additionally, HNB is the best filtering with confidence of 100% for the Apriori and 93.38% for the HotSpot. This may have occurred since HNB performs well on classes with a low distribution.

Secondly, we found that the Apriori can build high confidence rules (93.97% to 100%) and with minimum support of 30% to 45%, whereas the Hotspot algorithm can generate the rules with minimum confidence of 65% up to 93% and minimum support of 30% to 80%. These results are similar to those of Sharma and Om [31]. They reported that the Apriori algorithm can produce confidence up to 100%. However, the minimum support defined for a generated rule was 10%. This may be because the Hotspot algorithm employs a tree structure resulting in a high probability of rules in minority class with low minimum support.

Finally, the HotSpot algorithm can produce more rules with a significant minimum support than the Apriori algorithm. This may be due to the fact that the Apriori algorithm produces rules according to itemsets based on support-confidence, while the HotSport algorithm produces rules from the leaves of a tree structure. This finding is consistent with those of Sharma and Om [31]. Their results showed that the Apriori algorithm can only provide oral cancer survival rules with low minimum support (10%).

The strength of this study is its large sample size. This study gives insight into the usefulness of radiation therapy, as well as medical treatments. Some weaknesses of the study include the retrospective nature of the study and potential inaccuracies of the SEER database. For the further work, we plan to build a more comprehensive Apriori algorithm to improve rule confidence with a high support factor.

5. Acknowledgements

We thank the SEER program for making data available for this work. Also thanks to the Faculty of Informatics, Mahasarakham University for funding this research.

6. References

[1] Cancer Network. Colon, rectal, and anal cancers. [Internet]. Connecticut: Cancer Network Home of the Journal Oncology; 2016 [cited 2016 April 22]. Available from: http://www.cancernetwork.com /cancer-management/colon-rectal-and-anal-cancers.

[2] Haggar F, Boushey R. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clin Colon Rectal Surg. 2009;22(4):191-7.

[3] Amarican cancer Society. Colorectal cancer : facts and figures 2014-2016. Atlanta, Georgia: Amarican cancer Society; 2016.

[4] Leelakusolvong S. Colorectal cancer: early detection can mean a cure [Internet]. The Nation; [Update 2014 October 28; cited 2016 Feb 28]. Available from: http://www.nationmultimedia.com/news/life/living_health/30246366.

[5] Information and Technology Division National Cancer Institute. Hospital based cancer registry annual report 2013. Bangkok: National Cancer Institute Department of Medicine services, Ministry of Public Health, Thailand; 2015.

[6] Khuhaprema T, Srivatanakul P. Colon and rectum cancer in Thailand: An overview 2008. Jpn J Clin Oncol. 2008;38(4):237-43.

[7] Rebecca S, Carol D, Ahmedin J. Colorectal cancer statistics. CA Cancer J Clin. 2014;64(2):104-17.

[8] American Cancer Society. Key statistics for colorectal cancer [Internet]. Atlanta: American Cancer Society; 2016 [cited 2016 Augus 16]. Available from: http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-key-statistics.

[9] Chen J, He H, Jin H, McAullay D, Williams G, Kelman C. Identifying risk groups associated with colorectal cancer. In: Williams GJ, Simoff SJ, editors. Data Mining, LNAI 3755. Heidelberg: Springer-Verlag Berlin Heidelberg; 2006. p. 260-72.

[10] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca JB, Jarke M, Zaniolo C, editors. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 94); 1994 Sep 12-15; Santiago, Chile. San Francisco: Morgan Kaufmann Publishers; 1994. p. 487-99.

[11] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record. 1993;22(2):207-16.

[12] Leung CK-S. Mining uncertain data. WIREs Data Mining and Knowledge Discovery. 2011;1(4):316-29.

[13] Vinnakota S, Lam NS. Socioeconomic inequality of cancer mortality in the united states: A spatial data mining approach. Int J Health Geogr. 2006;5(9): 1-12.

[14] Agrawal A, Choudhary A. Identifying hotspots in lung cancer data using association rule mining. In: Spiliopoulou M, Wang H, Cook D, Pei J, Wang W, Zaïane O, et al., editors. Proceeding of the 11th IEEE International Conference on Data Mining Workshops; 2011 Dec 11; Vancouver, Canada. New Jersey: IEEE Computer Society; 2011. p. 995-1002.

[15] Cuet CV, Szeja S, Wertheim BC, Ong ES. Disparities in treatment and survival of white and native american patients with colorectal cancer: a SEER analysis. J Am Coll Surg. 2011;213(4):469-74.

[16] O'Connor ES, Greenblatt DY, Loconte NK, Gangnon RE, Liou J-I, Heise CP, Smith MA. Adjuvant chemotherapy for stage Ii colon cancer with poor prognostic features. J Clin Oncol. 2011;29(25): 3381-8.

[17] Fathy SK. A predication survival model for colorectal cancer. In: Zemliak A, Mastorakis N, editors. Proceeding of the American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications; 2011 Jan 29-31; Puerto Morelos, Mexico. Stevens Point: World Scientific and Engineering Academy and Society; 2011. p. 36-42.

[18] Fawzy A, Mokhtar HMO, Hegazy O. Outliers detection and classification in wireless sensor networks. Egypt Informat J. 2013; 14(2):157-64.

[19] Tallon-Ballesteros AJ, Riquelme JC. Deleting or keeping outliers for classifier training?. Proceeding of the 6th World Congress on Nature and Biologically Inspired Computing; 2014 Jul 30 - Aug 1; Porto, Portugal. New Jersey: IEEE; 2014. p. 281-6.

[20] Upadhyaya S, Singh K. Classification based outlier detection techniques. Int J Comput Trends Tech. 2012;3(2):294-8.
[21] Chenaoua K, Kurugollu F, Bouridane A. Data cleaning and outlier removal: application in human skin detection. Proceeding of 5th European Workshop on Visual Information Processing; 2014 Dec 10-12; Paris, France. New Jersey: IEEE; 2014. p. 23-28.
[22] Thongkam J, Xu G, Zhang Y, Huang F. Support vector machines for outlier detection in cancers survivability prediction. In: Ishikawa Y, He J, Xu G, Shi Y, Huang G, Pang C, et al., editors. Proceeding of International Workshop on Health Data Management: APWeb'08; 2008 April 26-28; Shenyang, China. Berlin: Springer; 2008. p. 99-109.
[23] Barbara D, Couto J, Jajodia S, Wu N. Detecting novel network intrusions using bayes estimators. In: Kumar V, Grossman R, editors. Proceeding of the 1st SIAM International Conference on Data Mining; 2001 April 5-7; Chicago, USA. Philadelphia: Society for Industrial and Applied Mathematics; 2001. p. 1-17.
[24] Farida DM, Zhanga L, Rahmanb CM, Hossaina MA, Strachana R. Hybrid decision tree and Naive Bayes classifiers for multi-class classification tasks. Expert Syst Appl. 2014 ;41(4):1937-46.
[25] Paris IHM, Affendey LS, Mustapha N. Improving academic performance prediction using voting technique in data mining. World Acad Sci Eng Tech. 2010;4(2):306-9.
[26] Kohavi R. Scaling up the accuracy of Naïve Bayes classifiers: A decision tree hybrid. In: Simoudis E, Han J, Fayyad U, editors. Proceeding of 2nd International Conference of Knowledge Discovery and Data mining; 1996 Aug 2-4; Portland, USA. California: AAAI Press. p. 202-7.
[27] Mahmood DY, Hussein MA. Analyzing NB, DT and NBtree intrusion detection algorithms. J Zankoy Sulaimani-Part A. 2014;16(1):87-94.
[28] Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987;27(3):221-34.
[29] Xiong H, Pandey G, Steinbach M, Kumar V. Enhancing data analysis with noise removal. IEEE Trans Knowl Data Eng. 2006;18(3):304-19.
[30] Gamberger D, Lavrac N, Groselj C. Experiments with noise filtering in a medical domain. In: Bratko I, Dzeroski S, editors. Proceeding of the 16th International Conference on Machine Learning. 1999 Jun 27-30; Bled, Slovenia. San Francisco: Morgan Kaufmann Publishers; 1999. p. 143-51.
[31] Sharma N, Om H. Significant patterns for oral cancer detection: association rule on clinical examination and history data. Netw Model Anal Health Inform Bioinforma. 2014;3(1):1-13.
[32] Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F. An application of association rule mining to extract risk pattern for type 2 diabetes using teharan lipid and glucose study database. Int J Endocrinol Metab. 2015;13(2):1-13.
[33] Arikan U, Gurgen F. Discrimination ability of time-domain features and rules for arrhythmia classification. Math Comput Appl. 2012;17(2):111-20.
[34] The University of Waikato. Weka 3: data mining software in Java [Internet]. Hamilton, New Zealand: Machine Learning Group at the University of Waikato; 2016 [cited on 2016 August 16]. Available from: http://www.cs.waikato.ac.nz/ml/.