Data Classification Based on Feature Selection with Association Rule Mining

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan,
Kittisak Kerdprasop, and Nittaya Kerdprasop

Abstract— **The aim of this paper is to study the problem of finding the optimal set of features that influence to the class discovery and to propose a novel method for feature selection. Currently, with the advancement of computer and internet technologies, new data is tremendously increasing every day causing a big data problem. This situation has made automatic data classification a difficult task. Reducing data dimensions to the minimal set of features is one solution to such problem. Therefore, this paper intends to solve the feature selection by proposing a method based on association analysis for analyzing features most influencing the class attribute. Experimental results confirm efficacy of our proposed method.**

Index Terms—**features selection, association rule mining, data classification**

I. Introduction

Current technologies have extensive role in the daily life of people such as Facebook, Twitter, and online shopping, resulting in the continuously generating of new data every day, which is both useful and useless. It is difficult to analyze and build meaningful models from these huge amount of data because it takes so much time to process that the analysis results cannot be obtained on time.

Data quality is important for the classification process in such a way that low quality data can degrade the performance of the model construction. This low performance problem is due to the fact that there are too many irrelevant features that do not contribute to the final model but they have to be evaluated during the model construction process. So, many researchers try to solve this problem by proposing several techniques to filter out useless features. These techniques can be generally divided into 2 groups: feature selection and feature extraction. Feature selection is the process of evaluating and taking only potentially useful subset of features without changing their original forms. Feature extraction, on the contrary, reducing number of features by transforming them to a more discriminative subspace. Conventional feature extraction technique used in many applications is principal component analysis.

Both feature selection and feature extraction are popular techniques applied to solve the classification problem from data with too many features. These data reduction techniques use some measures to calculate weight and then choosing features ordered by the weight [1, 2]. There are many researches trying to create new measure to calculate weight for reducing number of features and at the same time increasing accuracy of the final model. There is a research work using association rule mining technique to calculate weight, but the proposed process is quite complex [3].

Association rule mining is a well-known technique in data mining. It is the induction of relationships of events or objects and these relationships can be represented as rules for the ease of understanding and the convenience for applying the rules to predict the occurrence of an event or object in the future. There are many efficient techniques for performing the association rule mining, such as Apriori [4], Eclat [5], and FP-growth [6].

This research aims at proposing an efficient algorithm for data classification integrated with feature selection process based on the association rule mining using Aprori algorithm to generate rules that have high impact on the class attribute. We focus the impact through the high confidence of association rules to ensure feature appearance in the final model.

The contributions of this paper are as follows:

- With the proposed method, association rule mining can be applied for feature selection.
- The proposed method can reduce the number of features and at the same time can increase the model accuracy.

II. Materials and methods

A. Feature Selection

Feature selection is the process of calculating importance of each feature and then selecting the most discriminative subset of features. In data classification, some data (such as genetic data sets) may have thousands of features. Building a classification model from such high dimensional data may

Manuscript received November 22, 2016; revised January 10, 2017. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

N. Kaoungku is with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nuntawut@sut.ac.th).

K. Suksut is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, NakhonRatchasima, Thailand (e-mail: mikaiterng@gmail.com).

R. Chanklan is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, NakhonRatchasima, Thailand (e-mail: arc_angle@hotmail.com).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand. (e-mail: kittisakThailand@gmail.com).

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand. (e-mail: nittaya@sut.ac.th).

result in a low performance learning process. Therefore, many researchers try to solve this problem using feature selection technique. Feature selection is the removing of features that are not important and keeping only the important ones. Feature selection can be divided further into two classes of selection techniques [7]:

- Filter method. It is feature selection using the calculation of weight, which may be the relationship between features and class, and then choosing features having weight higher than some specific threshold. The algorithms in this category include CfsSubsetEval [8], Information Gain, and Chi-Square [9]. Fig 1 show the process of filter method.
- Wrapper method. It is the feature subset selection in which the subset generation and the learning algorithm are wrapped inside the same module. The subset selection steps can be iterative for the best model creation yielding high classification accuracy. Fig 2 shows the process of wrapper method [10, 11].
- Embedded method. It is the feature selection that is part of classification. It is advantage combination for both filter and wrapper methods by selecting features together with creating model. Fig 3 shows the process of embedded method [11].

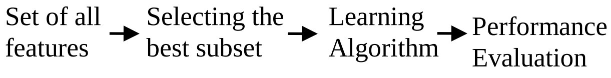


Fig. 1. Process of filter method [7].

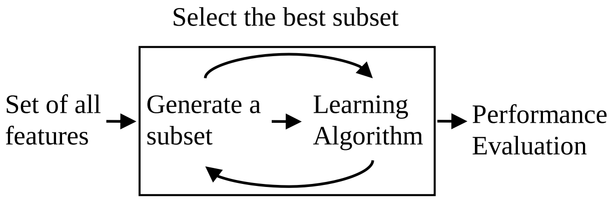


Fig. 2. Process of wrapper method [7].

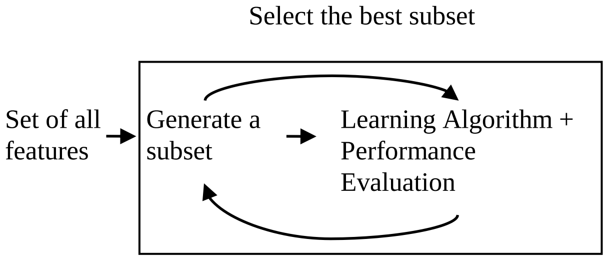


Fig. 3. Process of embedded method [7].

B. Association Rule Mining.

Association rule mining is the search for relationship of the event or frequent pattern that has the potential to be applied in the analysis or predicting the future events. This relationship is in the form *"IF condition Then consequence"* [12]. To limit the search space, the discovered relationship has to satisfy these two criteria:

- Support. It is the frequency of the occurring event. The occurrence frequency can be computed as the probability that two events (A and B) occur in the same transaction (equation 1). The minimum threshold of support value is normally specified by users.

$$Support(A \rightarrow B) = P(A \wedge B) \quad (1)$$

- Confidence. It is the proportion of frequency of co-occurring events (A and B) to the frequency of antecedent event (A). The computation is in equation 2. The minimum confidence is the threshold used to screen only interesting relationships.

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A)} \quad (2)$$

C. Apriori Algorithm

Apriori is a well-known algorithm for association rule mining. The algorithm finds frequent itemsets from database, but reduces the unnecessary search by deleting the itemset that its frequency is lower than minimum support [4]. Fig. 4 show the Apriori algorithm, which consists of five step.

Step 1: scan database to count items and calculate support, and then generate 1-itemset frequent pattern (L_1) containing only items having support value higher than the minimum support.

Step 2: from line 1 to 2, generate candidate itemset (C_{k+1} with k=1, 2, 3…, n) from frequent itemset (L_k with k=1, 2, 3…, n).

Step 3: from line 3 to 4, scan database to count support values of all candidates in C_{k+1}.

Step 4: from line 5, generate frequent itemset L_{k+1} from C_{k+1} with more than minimum support.

Step 5: repeat steps 2 to 5 until L_k is empty. The algorithm then returns all frequent itemsets of the given database. After that the association rules, having confidence higher than the minimum threshold, can be generated from these frequent itemsets.

```
Algorithm Apriori

      $C_k$: Candidate itemset of size k
      $L_k$ : frequent itemset of size k
      $L_1$ = {frequent items};

1.    for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
2.       $C_{k+1}$ = candidates generated from $L_k$;
3.       for each transaction $t$ in database do
4.          increment the count of all candidates in $C_{k+1}$
            that are contained in $t$
5.       $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
6.    end
7.    return $\cup_k L_k$;
```

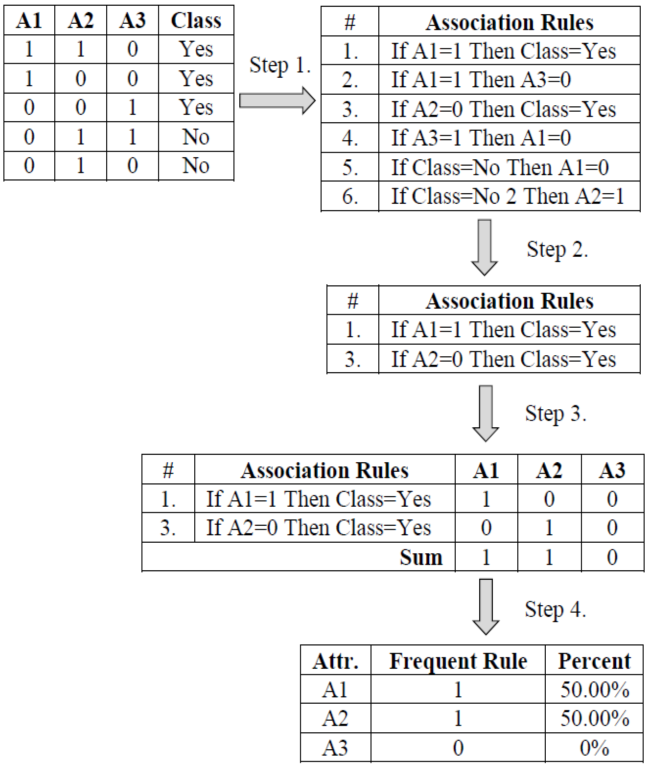
Fig. 4. Apriori algorithm [4].

III. Proposed Work

In this section, we present the proposed process of data classification based on feature selection with association rule mining. The intuitive idea is that we use association rule mining algorithm to build rules having the class attribute as their consequence part. The eligible rules are the ones with high confidence values. This is because these rules are frequent patterns inducible from the training data set and they are supposed to contain features that are important to class attribute.

Figure 5 shows running example for capturing our intuitive idea. At the first step, applying Apriori algorithm to find frequent patterns. Then build association rules from these frequent patterns. In Fig. 5, we set minimum support parameter to be 0.1, minimum confidence to be 0.9, and maximum number of conditional attribute to be 1. In Step 2, we then keep only rules that have class attribute (class = yes) as the rule consequence. In Step 3, we count the number of features from the selected rules (rules 1 and 3) and calculate the percentage of attribute appearance frequency (equation 3). Finally, we have features with calculated frequency as: A1 (50.00%), A2 (50.00%), and A3 (0%). These features can be ranked in the descending order according to their importance as either <A1, A2, A3>, or <A2, A1, A3>.

$$FrequentFeature(A) = \frac{AppearFrequency(A)}{\#Rules} \qquad (3)$$



A1	A2	A3	Class
1	1	0	Yes
1	0	0	Yes
0	0	1	Yes
0	1	1	No
0	1	0	No

#	Association Rules
1.	If A1=1 Then Class=Yes
2.	If A1=1 Then A3=0
3.	If A2=0 Then Class=Yes
4.	If A3=1 Then A1=0
5.	If Class=No Then A1=0
6.	If Class=No 2 Then A2=1

#	Association Rules
1.	If A1=1 Then Class=Yes
3.	If A2=0 Then Class=Yes

#	Association Rules	A1	A2	A3
1.	If A1=1 Then Class=Yes	1	0	0
3.	If A2=0 Then Class=Yes	0	1	0
	Sum	1	1	0

Attr.	Frequent Rule	Percent
A1	1	50.00%
A2	1	50.00%
A3	0	0%

Fig. 5. Running example of feature selection based on association rule mining.

The algorithm for feature selection based on association rule mining for classification is given in Figure 5. Our algorithm consists of three phases.

The first phase is at line 1. This phase is for conventional association rule mining with Apriori algorithm from training data set (*D*). This phase requires three parameters, which are minimum support threshold (*minsup*), minimum confidence threshold (*minconf*), maximum number of attributes that can be appeared at the conditional (or antecedent) of the association rules (*maxlen*), and minimum frequency of features or attributes that appeared in association rules (*minfrequent*). This last threshold is for discarding features with low importance.

Phase 2 is the part from lines 2 to 6. This phase is the rule pruning, which is the deletion of rules that have attributes in their consequence part disagree with the specified subset of class attribute (*C*). This phase is for selecting from association rules the frequent patterns containing both predictive features and class attribute.

Phase 3 is the part from lines 7 to 14. The operation of this phase is to count the frequency of features appeared in association rules that are obtained from phase 2. We design these steps to iterate over each attribute and count the appearance frequency of attributes that appear in the conditional part of the association rules.

```
Algorithm  Feature Selection with Apriori Algorithm

//Input:   D, training data set.
           minsup, minimum support threshold.
           minconf, minimum confidence threshold.
           maxlen, maximum number of conditional
                  attributes.
           C, class attribute.
           minfrequent, minimum frequency of attributes
                  in set of association rules.

//Output:  F, a set of frequent features.

1.   R = Apriori(D, minsup, minconf, maxlen)
2.   For each rule r ∈ R do
3.     If consequence(r) != C Then
4.       delete r  from R
5.     End If
6.   End For
7.   For each attribute Attr from D do
8.     For each rule r ∈ R do
9.       If condition(r) = $Attr_i$ Then
10.        count_$Attr_i$ ++
11.      End If
12.    End For
13.    add $Attr_i$ and count_$Attr_i$ to F
14.  End For
15.  For each feature f ∈ F do
16.    If  FrequentFeature(f)  < minfrequent Then
17.      delete f  from F
18.    End If
19.  End For
20.  Return F
```

Fig. 6. Algorithm to select feature based on association rule mining

Phase 4 is the part from line 15 to 19. This phase is the features selection from subset of frequent features of rules (*F*) by deleting features that have percentage of frequency appearance in the set of association rules lower than the specified minimum frequency threshold. Finally, the algorithm returns the subset of features that has been considered high importance to class attribute prediction based on the analysis of their appearance in the set of association rules induced from the training data set.

IV. Experimental results

The proposed feature selection method has been experimented with real data from the UCI Machine Learning Repository. Table 1 show details of the nine data sets used in our experimentation. Each of these datasets has been divided into training dataset (70%) and test dataset (30%). We use the C4.5 algorithm for classification because of its popularity in many application areas ranging from scientific to business industries.

To run our feature selection algorithm, we set parameter of minimum support to be 0.1, minimum confidence to be 0.9, and maximum number of conditional attributes in the induced association rules to be 1. In case of such parameter setting generates an empty rule, we can reduce the minimum confidence or increase the maximum number of conditional attributes in the association rules.

The performance of our proposed feature selection method has been compared with the CfsSubsetEval, Gain Ratio, and Information Gain algorithms. The performance metrics are number of selected features (the lower is the better) and accuracy of the classification algorithm (the higher is the better). This work has been implemented with both RStudio and WEKA. We run our experiments on a core i5/2.30 GHZ computer with 4 GB of RAM.

Table 2 and Fig 7 show comparative results of classification accuracy after applying the four feature selection methods. It can be seen that the feature selection algorithm proposed in this research work can improve the performance of accuracy on Ecoli, Breast Cancer, Heart, Zoo, and Hepatitis data sets when compared to raw data set with no feature selection method and other feature selection algorithms. But the performance of our algorithm on the Wine and Vote data sets shows lower accuracy than some algorithms.

Table 3 and Fig 8 show comparative results of number of features obtained from the four feature selection algorithms. It can be seen that our feature selection algorithm can reduce the number of features on the Ecoli, Diabetes, Breast Cancer, Wine, Zoo, and Horse colic data sets, comparative to other feature selection methods. But the number of features selected from our algorithm on the Heart, Vote, and Hepatitis data sets shows higher number of features than some algorithms.

TABLE I
Details of Datasets

Datasets	**# Instances**	**# Attributes**
Ecoli	336	8
Diabetes	768	9
Breast Cancer	286	10
Heart	303	14
Wine	178	14
Vote	435	17
Zoo	101	18
Hepatitis	155	20
Horse colic	368	23

TABLE II
Comparative Results of Accuracy by Four Feature Selection Algorithms.

Datasets	**Algorithms**				
	Raw	**A**	**C**	**G**	**I**
Ecoli	79.79	**80.85**	79.79	69.15	79.79
Diabetes	**80.34**	**80.34**	**80.34**	**80.34**	**80.34**
Breast Cancer	71.25	**73.75**	**73.75**	**73.75**	**73.75**
Heart	77.38	**80.95**	72.62	72.62	72.62
Wine	**91.11**	88.89	**91.11**	**91.11**	86.67
Vote	**96.88**	96.09	94.53	96.09	96.09
Zoo	77.27	**90.91**	81.82	81.82	81.82
Hepatitis	71.79	**76.92**	**76.92**	**76.92**	**76.92**
Horse Colic	**86.79**	**86.79**	**86.79**	**86.79**	**86.79**

A = the proposed feature selection based on association rule mining,
C = CfsSubsetEval, G = Gain Ratio; I = Information Gain.

TABLE III
Comparative Results of Number of Features by Four Feature Selection Algorithms.

Datasets	**Algorithms**				
	Raw	**A**	**C**	**G**	**I**
Ecoli	7	**4**	6	**4**	5
Diabetes	9	**6**	**6**	**6**	**6**
Breast Cancer	9	**4**	**4**	5	5
Heart	13	8	7	**6**	7
Wine	13	**7**	11	11	9
Vote	16	9	**5**	8	8
Zoo	17	**7**	12	11	10
Hepatitis	19	12	10	12	**9**
Horse Colic	22	**5**	**5**	10	10

A = the proposed feature selection based on association rule mining,
C = CfsSubsetEval, G = Gain Ratio; I = Information Gain.

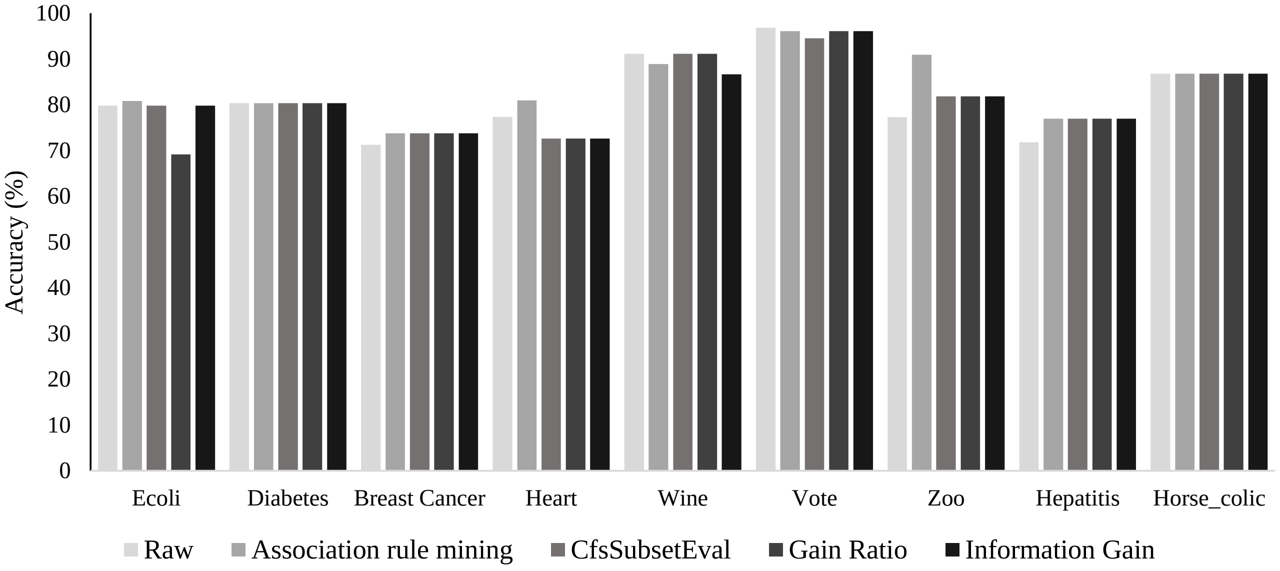


Fig. 7. The accuracy comparison of the four feature selection algorithms.

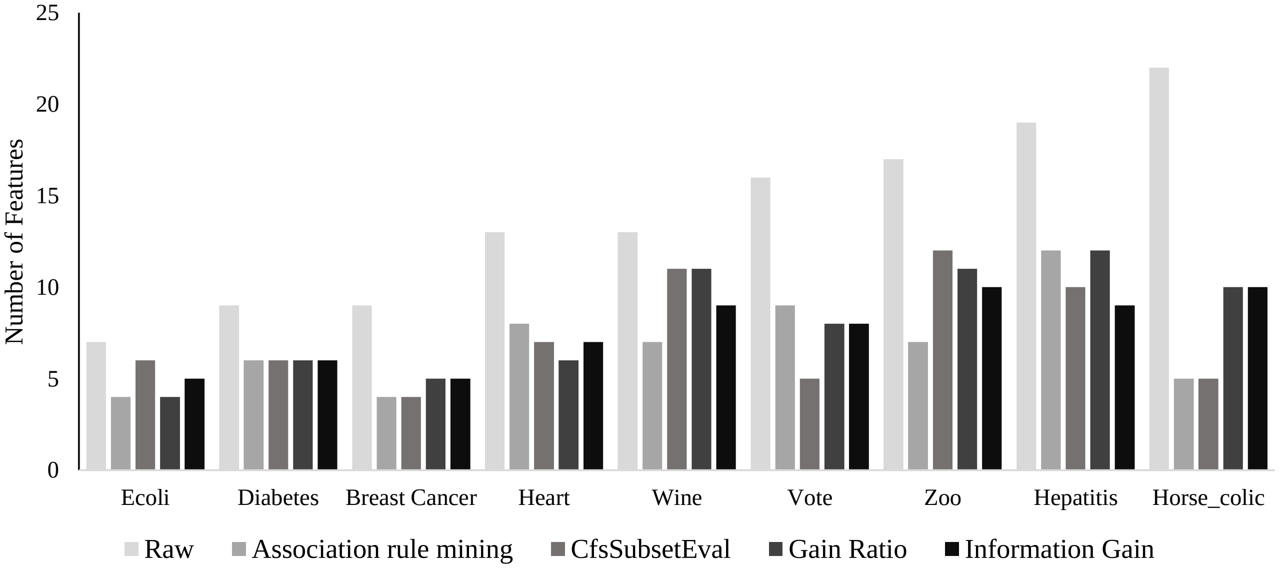


Fig. 8. The number of features comparison of the four feature selection algorithms.

V. Conclusion

This research aims at studying the data classification problem over high dimensional data based on the feature selection with association rule mining. The problem of data classification with many features is that the classification model may yield low accuracy and the model building phase consumes so many computation resources as a result of the existence of useless features in the data. Thus, we propose to use the association rule mining to generate rules that have rule consequence as class attribute. The reason is that these rules can shows frequent features that have high impact to the class attribute. We present in this paper the algorithm to consider frequent features from the induced association rules. Our algorithm also prunes features appearing less frequent as compared to the other discovered features.

From the experimental results, it has been revealed that the proposed algorithm can reduce the number of features, and at the same time it can also increase the accuracy in data classification. However, the feature selection of our proposed method over some datasets show lower accuracy than other feature selection methods, but the trade-off is the fewer number of features.

References

[1] I. Guyon, “Practical feature selection: from correlation to causality,” *NATO Science for Peace and Security*, vol. 19, pp. 27-43, 2008.

[2] H. G. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem,” in *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 121-129, 1994.

[3] J. Xie, J. Wu, and Q. Qian, “Feature selection algorithm based on association rules mining method”. in *Proceedings of the ACIS International Conference on Computer and Information Science*, pp. 357-362, 2009.

[4] R. Agrawal, and R. Srikant, “Fast algorithms for mining association rules”, in *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, 1994.

[5] J. M. Zaki, "Scalable algorithms for association mining", *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 3, pp. 372–390, 2000.

[6] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation”, *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1-12. 2000.

[7] M. Hilario, and A. Kalousis, “Approaches to dimensionality reduction in proteomic biomarker studies”, *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 102-118, 2008.

[8] Z. N. Hamilton, “Correlation-based feature subset selection for machine learning”, *PhD Thesis*, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 1999.

[9] X. Jin, A. Xu, R. Bie, and P. Guo, “Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles”, in *Proceedings of the International Workshop on Data Mining for Biomedical Applications*, pp. 106-115, 2006.

[10] L. Yu, and H. Liu, “Efficient feature selection via analysis of relevance and redundancy”, *The Journal of Machine Learning Research*, vol. 5, pp. 1205-1224, 2004.

[11] Y. Saeys, P. Rouzé, and Y. Van de Peer, “In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists”, *Bioinformatics*, vol. 23, no. 4, pp. 414-420, 2007.

[12] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases”, *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.