GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Proposed method for dimensionality reduction based on framework in gene expression domain

D.C. Macedo[1], E.C.M. Ishikawa[2], C.B. Santos[3], S.N. Matos[4], H.B. Borges[5] and A.C. Francisco[6]

Departamento de Engenharia de Produção,
Universidade Tecnológica Federal do Paraná, Ponta Grossa, PR, Brasil

Corresponding author: D.C. Macedo
E-mail: dayanamacedo@yahoo.com.br

Genet. Mol. Res. 13 (4): 10582-10591 (2014)
Received October 30, 2014
Accepted December 1, 2014
Published December 12, 2014
DOI http://dx.doi.org/10.4238/2014.December.12.21

ABSTRACT. The excessive use of attributes may affect the search for patterns and extraction of useful knowledge, because they harm the learning performance of algorithms in both speed and success rate. The use of dimensionality reduction methods is therefore an important alternative; however, these methods do not deal with the reduction of attributes in a specific area. This article presents a method based on framework concepts of domain for reducing attributes in a domain. The input method is a set of databases related to a domain, and the main process is the identification of common and variable attributes, plus the reduction of attributes in the original database. The proposed method was applied in the gene expression domain, using databases. The method can be used to analyze the most relevant attributes in a specific domain, granting greater confidence for models created for the application of a data mining task, thus, a previously known method in data mining. Attribute selection was also applied in the three databases for the comparison of the results. Analyses of the results using the criterion of cross-validation revealed that the employment of the methods resulted in the improvement of success rates compared to the databases

containing the full range of attributes.

Key words: Dimensionality reduction; Framework; Attribute selection

INTRODUCTION

The generation of patterns occurs when there is a need to develop a set of combinations of attributes capable of fulfilling the specific necessities of an application (Prahalad and Krishnan, 2008). The significant number of attributes usually present in a database may harm the search for patterns and the extraction of useful knowledge. For instance, in a customer database, an attribute may be important while others not. As a result, there is a need to select the most relevant attributes (Romdhane et al., 2010). Accordingly, it is important to reduce the amount of data and existing attributes to stop them from hindering the learning process.

According to Kira and Rendell (1992), redundant attributes harm the learning performance of algorithms in speed (due to data dimensionality) as well as in the success rate (due to the presence of redundant information that may confuse the algorithm, instead of helping with the search for a correct knowledge model). In data mining, some techniques are used for the dimensionality reduction of attributes in databases; especially the attribute selection method. According to Kira and Rendell (1992), the goal of attribute selection is to select, in a single database, a subset of relevant attributes to improve the importance of the learning process and ensure data quality.

This paper proposes a method capable of identifying the most relevant attributes in n databases of a domain. A domain is defined as a set of characteristics that describe a group of problems where a particular application is used to propose a solution (Clements and Northrop, 2002; Pohl et al., 2005).

This method was created through domain Framework concepts (Froehlich et al., 2000), which refine the attributes.

Johnson and Foote (1988) define Framework from a structural point of view, as being a "set of abstract and concrete classes that form the abstract project for a group of related problems". From the point of view of purpose, framework is defined as a structure of an application that is instantiated by the developer of applications (Johnson, 1997). Framework allows for reuse of code, project or analyses. The reuse of analyses is obtained because it describes the objects, their relationships and the way by which big problems are modularized (Budd, 2002). The reuse of projects occurs when framework contains abstract algorithms and the definition of their interfaces, such as the obstacles of an implementation.

The use of framework provides benefits to the development of information systems (Kubo and Tori, 2006), such as the following: i) There is an increase in speed and a decrease in time, because it uses pre-fabricated and pre-tested components. There is no need for the developer to find out new classes and build interfaces. There is the relevance of rewriting the behavior of specific methods of certain classes. Program structure and execution flow are already specified. ii) Framework allows the reuse of code and project by inheritance and/or polymorphism. iii) There is a reduction of maintenance costs and needs when Framework is the majority of the code required by applications. Due to inheritance, when an error in Framework is corrected, or when there is the addition of a characteristic, the benefits are immediately extended to the new class. iv) It is possible to develop ever more complex and powerful applications from existing frameworks.

The existing approaches in the literature are intended for assisting the development of domain frameworks. This study was based on the method of Ben-Abdallah (2004) because it establishes a set of relations and rules that enable the analysis of the example-applications in a domain during the creation of a model. It also has the capacity of defining concepts of equivalence and generalization. This method focuses on the thematic of class diagrams, using the comparison criteria of classes, attributes and operations. Since it has a specific comparison for attributes, this method was used as a basis for the development of the proposed method for dimensionality reduction, attribute selection.

This method focuses on the thematic of class diagrams, using the comparison criteria of classes, attributes and operations.

Thus, attribute selection and the proposed method were used in databases in the gene expression domain to provide comparisons between them. These methods were applied in three databases: DLBCL, DLBCL - Tumor and AML/ALL in the gene expression domain.

The attribute selection method was carried out by the Filter and Wrapper approaches according to Liu and Yu (2005) in the Waikato Environment for Knowledge Analysis data mining environment (Weka, 2011). In the filter approach, two algorithms were used: correlation-based feature selection (Hall, 1999) and consistency subset evaluator (Tan et al., 2008). For the wrapper approach, the classification algorithms used were: naïve Bayes, J48, support vector machines (SVM) and k-NN (for k = 1, k = 3, k = 5, and k = 7) (Borges and Nievola, 2012). The method was used through an algorithm for the identification of the most applicable attributes, and after the generation of a new database created from framework concepts, the Weka environment was used. Success rate was chosen as the evaluation criterion, obtained by cross-validation to see whether the use of this method contributes significantly to success rates.

This article is organized as follows: Material and Methods describes the proposed method and experiments. Then results are presented, using attribute selection and framework concepts in the gene expression domain. Finally, the last section presents the Discussion of this study.

MATERIAL AND METHODS

The development of the study followed the three main steps of the knowledge discovery in database process: pre-processing, data mining and post-processing. Part of the execution of the experiments was conducted in the Weka (Waikato Environment for Knowledge Analysis) software.

Dimensionality reduction method based on framework (DRM-F)

The proposed method, DRM-F, is composed of the three steps presented in Figure 1.

The first step relates to the identification of the domain in which some of the tasks of data mining shall be achieved. A domain is defined as a set of features that describe a group of problems where a certain application intends to propose a solution. In the realm of biology, there is the study of genes to determine whether or not a person is susceptible to a disease.

The second step prepares the database and it is composed of three sub-steps: pre-process procedure to ensure its integrity, attributes transformation and apply domain framework concept. The application of the framework concept for attribute selection involves the analysis of all the attributes of all the databases to determine if they are common or specific ones. The

next step in the proposed algorithm is to evaluate the databases by mining algorithms, to analyze what resulted from the use of the framework concepts in attribute selection.

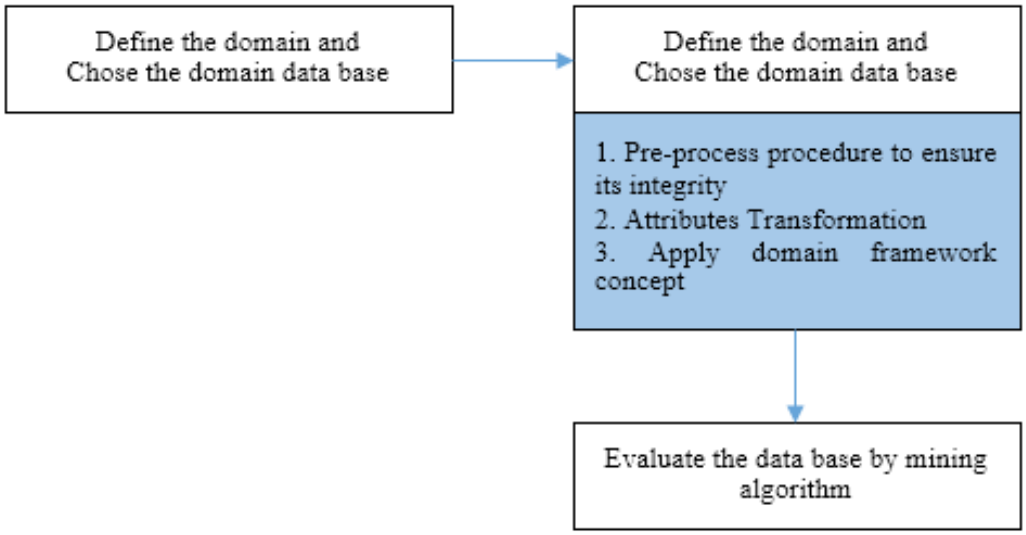

Figure 1. Steps of the proposed method (DRM-F).

Description of the data sets

For the application of the proposed method, DRM-F, it is necessary to use at least three databases of a domain. Data sets were selected, gene expression one, obtained by the microarray technique. The bases are called DLBCL, DLBCL - tumor and DLBCL ALL/AML. The DLBCL base has gene expression data on Diffuse Lymphoma cancer of large B cells. The base was obtained by means of the use of micro arrangement techniques. The type of cancer studied by the authors is the subtype of the non-Hodgkin lymphoma, which commonly constitutes the group of cancers of malignant tumors. It was possible to identify two distinct forms of cells of diffuse lymphoma of large B cell, which had gene expression patterns indicated by two different stages of B cell differentiation. Considering that a type of gene expression characterizes the germinal center of B cell and the second type of gene expression is the activation of the B cell. Fluorescent images of micro hybrid arrangement were obtained using the GenePix 4000 microarray scanner. It is important to emphasize that in this base there are 4026 attributes (genes) and 47 examples, of which 24 belong to the group of germinal center of B cell, while 23 belong to the group of activation of the B cell. Each example is described by the 4026 attributes (genes). The DLBCL base - Tumor was obtained by means of the micro array called Hu6800 Affymetrix. This set consists of two types of lymphoma: The diffuse lymphoma of large B cells and the follicular lymphoma. There are in this base 7129 attributes (genes) containing 77 specimens, of which 58 belong to the group of diffuse lymphoma of large B cells and 19 belonging to the group of follicular lymphoma. Finally the base ALL/AML contains the analysis of two types of acute leukemia, they are: acute lymphoblastic leukemia and acute myeloid leukemia. Also in this base the used micro-

array technique is the Hu6800 Affymetrix. The database contains samples from 72 patients, 47 of whom refer to the acute lymphoblastic leukemia type and 25 of acute myeloid leukemia. There are in total 7129 attributes (genes) on this base.

In the domain of gene expression data, three sets were extracted from the Kent Ridge Bio-Medical Dataset Repository (2013), which were about the study of lymphoma and leukemia. These sets were already used by Borges and Nievola (2012) in their studies. Table 1 presents the characteristics of these databases.

Table 1. Data bases characteristics.

Domain gene expression	Data base	Amount of attributes	Amount of samples
	DLBCL (Alizadeh et al., 2000)	4026	47
	DLBCL-Tumor (Shipp et al., 2002)	7129	77
	AML-ALL (Golub et al., 1999)	7129	72

RESULTS

This section presents the results obtained by the dimensionality reduction methods through attribute selection and the proposed method, DRM-F, based on framework concepts. It is important to note that the attribute selection results presented, concerning the gene expression domain by attribute selection, are from the research of Borges and Nievola (2012). The three chosen databases were subjected to the seven classifiers. In the tables containing the results, the asterisk indicates that a result is significantly worse than the result of the standard algorithm (naïve Bayes), and that the result in bold indicates a slight but significant improvement compared to the standard algorithm. Also mentioned are the success rates obtained by the methods, seeking to identify the best results in each domain, in the respective databases, to evaluate the applicability of the methods.

General comparison of the dimensionality reduction methods

This section aims to compare the methods used in this research to determine which is better for dimensionality reduction in the databases. Hence, it includes a comparison of the two approaches used in the attribute selection method (filter and wrapper) and the DRM-F method adapted from Ben-Abdallah et al. (2004), for the identification of common and specific attributes in the segments.

Following is presented a comparative analysis of the concepts used for dimensionality reduction in the gene expression domain.

General comparison in the gene expression domain

A comparison of the results from the application of the dimensionality reduction methods was also made for the gene expression domain. Table 2 provides the results for the DLBCL database.

To compare the methods, it is necessary to determine the results with the dimensionality reduction methods for the original database with all the attributes, so that the application of the methods can be justified. Thus, on average, the classification algorithms in the database with all the attributes achieved an 81.86% success rate. The methods used

in attribute selection showed an average success rate of 95%, compared to 79.33% with the proposed method.

The attribute selection method gave a better result compared to the original database, and over the proposed method. The proposed method had a result close to that of the original database, which was lower by 2.33%. However, this result does not render the method unusable. Because its result was over 80%, the method is still applicable. Particular attention is drawn to the fact that the proposed method takes into account the meaning of each attribute in the context of the domain, while selecting attributes.

Concerning attribute selection, the wrapper approach performed better than the filter approach. The average value for the performance of the classification algorithms was 97.35% in the wrapper approach and 93.81% in the filter approach. For the proposed method, the results identified in the DLBCL database demonstrated an average performance of 79% for the common attributes and 79.33% for the specific attributes.

From the values above, the attribute selection method performed better than the proposed method, with a 95% success rate average. On the other hand, the methods applying framework concepts reached an average performance of 79.33%. The best dimensionality reduction method in the DLBCL database was with the wrapper approach, and the worst results were related to the specific attributes. In the original database with all the attributes, the best classification algorithm was naïve Bayes. For attribute selection with the Filter the best algorithms were naïve Bayes and SVM in the dependency measure, and 3-NN in the consistency measure. In the best reduction method, with use of the wrapper approach, the 7-NN classification algorithm stood out. In the proposed framework-based method. SVM once again stood out as the best algorithm for common attributes in the gene expression domain. For specific attributes, naïve Bayes performed the best.

Regarding the lowest performances for success rate, J48 was the worst algorithm for four times in the DLBCL base, 7-NN was the worst algorithm for the original database and specific attributes. Table 3 illustrates a comparison of the reduction methods for the DLBCL - Tumor database.

Analyzing the wrapper approach, all the algorithms were statistically better than the standard; only J48 showed lower performance. About the common and specific attributes, all were worse than the standard. The SVM algorithm only gave the best result when compared with the standard adopted. Table 3 shows the comparative analysis of reduction methods for the DLBCL Tumor database, where the wrapper approach had the best average reduction performance, with a 97.36% success rate. The wrapper approach was also superior to the filter approach in attribute selection. Regarding the comparative analysis with the original database with all the attributes, wrapper again gave the best performance, since the original database had an average success rate of 86.76%.

Analyzing the results of the proposed method, DRM-F, the common attributes had an 87.82% average success rate versus 85.30% for the specific ones. Thus, the proposed method was better for the common attributes. Among the dimensionality reduction methods, attribute selection stood out once more, yielding results superior to those of the proposed method. Notwithstanding, the proposed method showed a higher than 80% success rate, which is not a bad result. It should be taken into account that the search criterion for the proposed method was based on the identification of common attributes in the chosen and specific domain databases analyzed. The proposed method is intended to search for the equivalence and generalization of attributes in the domain studied. From the information above, the best classification algo-

rithm for the original database with all the attributes and the filter approach with dependency measure was SVM. In the consistency measure, it was naïve Bayes instead. For the wrapper approach, the best algorithms were naïve Bayes and 1- NN.

For the proposed method in the common attributes, the best performance was obtained with the SVM classifier, and 3-NN and 5-NN for the specific attributes. The J48 algorithm had the worst performance among all algorithms on four occasions in the original database with all attributes; filter approach and dependency measure for common and specific attributes. For the wrapper approach, 3-NN and 7-NN both performed worst. SVN showed inferior performance in the filter approach method with consistency measure for specific attributes. In relation to statistical analysis, all algorithms had better performance than that adopted as the standard.

Table 4 illustrates the results obtained by the reduction methods for the ALL/AML database. In the ALL/AML database, the best dimensionality reduction method was with the wrapper approach in attribute selection. The average success rate for this method was 98.82%. It is important to note that, for this database, the proposed method, DRM-F, showed a better performance compared to the original database. On average, DRM-F had a 87.97% success rate versus 86.48% obtained with the original database with all attributes.

An isolated analysis of the attribute selection method by means of the filter approach indicates that the use of the dependency measure gave superior results compared to the consistency measure. The dependency measure had an average success rate of 94.92 versus 89.38% with the consistency measure. With this information in the filter approach, the dependency measure had the best performance. The wrapper approach performed better compared to the filter approach, with 98.82% average success rate. Hence, it was the best attribute selection method for dimensionality reduction in this research.

Regarding the proposed method, DRM-F, the results for common attributes were also superior in performance to those for specific attributes, the same was found for the DLBCL database. For the common attributes, the average success rate of the classification algorithms was 88.46%, compared to the 87.47% average obtained for the specific attributes.

Comparing the attribute selection and the framework concept method, the former stood out once more with a superior performance, notwithstanding the fact that the framework concepts are designed to search for similarities between attributes in the context of the domain of choice. Attribute selection, on the other hand, a specific search logic was used that takes into consideration the correlations between the attributes concerning the meta attribute, or class attribute. According to the above, the best classification algorithms in the original database were naïve Bayes and SVM. For the filter approach in the dependency measure, the best performers were 5-NN and 7-NN, and naïve Bayes for the consistency measure. In the wrapper approach, Naïve Bayes was, once again, the algorithm with the best results, together with the k-NN algorithms, with a 100% success rate. The lowest performance in the original database with all attributes was found for the 7-NN algorithm. In the filter approach using the dependency measure, J48 performed worst, as well as in the wrapper approach and common attributes.

In the statistical analysis, it could be seen that all the classifier algorithms had a higher performance than the standard adopted for the datasetP, filter and wrapper approaches. But for the DRM-F method in relation to the common and specific attributes, this had a lower performance than the naive Bayes. The SVM algorithm had a higher value than the standard used for the specific attributes.

Table 2. Average of the classification algorithms in the Attribute Selection and DRM-F methods for the DLBCL base.

Subsets	Algorithms						
	Naïve Bayes	J48	SVM	1-NN	3-NN	5-NN	7-NN
All the Attributes Dataset	97.5 ± 7.9	77.0 ± 23.7	**98.0 ± 6.3**	75.5 ± 21.2	77.0 ± 17.5	75.0 ± 23.7	73.0 ± 18.7*
Consistency	100.0 ± 0.0	76.5 ± 30.0*	**100.0 ± 0.0**	98.0 ± 6.3	98.0 ± 6.3	96.0 ± 8.4	96.0 ± 8.4
Dependency	94.0 ± 9.7	89.0 ± 11.7*	89.5 ± 11.2*	93.5 ± 10.5	**96.0 ± 8.4**	93.5 ± 10.5	93.5 ± 10.5
Wrapper	98 ± 6.32	91.5 ± 11.07*	**98.0 ± 6.32**	98 ± 6.32	98 ± 6.32	98 ± 6.32	100 ± 0*
Common	97.5 ± 7.91	55.5 ± 20.74*	100 ± 0	75.5 ± 23.27	81.00 ± 15.06	68.5 ± 25.83	75.00 ± 18.26
Specifics	93.50 ± 14.15	73.22 ± 22.75	**96 ± 8.43**	78.50 ± 13.55	74.50 ± 19.50	73 ± 24.06	69.00 ± 21.19*

Asterisk indicates that a result is significantly worse than the result of the standard algorithm (naïve Bayes). Results in bold indicate a light but not significant improvement compared to the standard algorithm.

Table 3. Average of the classification algorithms in the Attribute Selection and DRM-F methods for the DLBCL - Tumor base.

Subsets	Algorithms						
	Naïve Bayes	J48	SVM	1-NN	3-NN	5-NN	7-NN
All the attributes	80.5 ± 10.7	72.5 ± 16.1	**96.1 ± 6.3**	84.1 ± 13.5	93.2 ± 9.8	89.8 ± 9.9	91.1 ± 8.5
Consistency	96.07 ± 3.34	**86.79 ± 12.26**	97.50 ± 5.27	93.57 ± 9.00	95.89 ± 6.63	97.32 ± 5.66	97.32 ± 5.66
Dependency	93.57 ± 6.80	90.89 ± 10.96	**76.61 ± 5.48**	91.96 ± 14.80	92.14 ± 9.00	89.64 ± 9.99	90.89 ± 8.64
Wrapper	100 ± 0	93.4 ± 11.36	98.75 ± 3.95	**100 ± 0**	97.3 ± 5.66	94.8 ± 6.71	97.3 ± 5.66
Common	83.04 ± 16.09	78.21 ± 16.43	**98.75 ± 3.95**	86.79 ± 10.75	90.54 ± 11.57	92.32 ± 8.87	85.07 ± 21.92
Specifics	80.71 ± 10.30	80 ± 15.49	80.00 ± 15.49	82.86 ± 12.46	**91.96 ± 9.68**	**91.96 ± 9.68**	89.64 ± 12.05

Results in bold indicate a light but not significant improvement compared to the standard algorithm.

Table 4. Average of the classification algorithms in the Attribute Selection and DRM-F methods for the ALL/AML base.

Subsets	Algorithms						
	Naïve Bayes	J48	SVM	1-NN	3-NN	5-NN	7-NN
All the attributes	98.6 ± 4.5	78.9 ± 15.6	98.6 ± 4.5	84.6 ± 18.1	83.4 ± 12.9	83.4 ± 12.9	**77.9 ± 11.3**
Consistency	96.07 ± 3.34	**86.79 ± 12.26**	97.50 ± 5.27	93.57 ± 9.00	95.89 ± 6.63	97.32 ± 5.66	97.32 ± 5.66
Dependency	93.57 ± 6.80	90.89 ± 10.96	**76.61 ± 5.48**	91.96 ± 14.80	92.14 ± 9.00	89.64 ± 9.99	90.89 ± 8.64
Wrapper	100.00 ± 0	94.64 ± 9.11	97.14 ± 6.02	**100 ± 0**	**100 ± 0**	**100 ± 0**	**100 ± 0**
Common	97.50 ± 5.27	78.93 ± 17.02*	98.57 ± 4.52	87.5 ± 13.81	85.89 ± 13.48	86.07 ± 9.55	84.82 ± 10.14
Specifics	95.71 ± 6.90	87.50 ± 13.81	**97.14 ± 6.03**	86.07 ± 17.53	84.82 ± 13.92	81.96 ± 11.56	79.11 ± 13.83*

Asterisk indicates that a result is significantly worse than the result of the standard algorithm (naïve Bayes). Results in bold indicate a light but not significant improvement compared to the standard algorithm.

DISCUSSION

The application of dimensionality reduction methods is important, since the pursuit of useful knowledge and patterns in databases exempts the presence of a significant number of attributes. The use of the database with all attributes can impair the performance of the learning process of the algorithms. Thus, there is a need to apply methods that ensure the quality of incoming data for the data mining phase. The amount of redundant information can confuse the algorithm instead of assisting in the search for a correct model of knowledge.

This paper compared two reduction methods: attribute selection method and framework-based method, called DRM-F. This comparison aimed at evaluating the proposed method and existing method in data mining, that is, attribute selection. DRM-F was adapted from Ben-Abdallah et al. (2004).

These two methods were applied in the field of gene expression, where three databases were used, namely DLBCL, DLBCL- Tumor on leukemia and ALL/AML containing lymphoma data. These sets have been used by Borges and Nievola (2012) in their studies. The three data sets were extracted from the Kent Ridge Bio-medical Dataset Repository.

Analyzing the results obtained using the cross-validation and holdout evaluation criteria, it was found that the use of the methods resulted in an improvement in the success rate values compared to the databases containing all the attributes of the gene expression domains.

In the gene expression domain, the best reduction method consisted of the wrapper approach for the three databases. Nevertheless, it is noteworthy that the proposed method showed >80% success rate, indicating that it cannot be considered a reduction method with poor performance. The results obtained were near those of the original database containing all attributes. Considering that the search criterion of the proposed method is based on the identification of common and specific attributes among databases analyzed in the chosen domain, the proposed method seeks to search for the equivalence and generalization of attributes in the domain studied.

Future efforts will seek to create an automated tool that is able to identify the common and specific attributes, as well as developing the DRM-F algorithm. For the gene expression databases, the attributes of the best reduction methods will be subjected to a specific analysis to determine the biological significance of the attributes, in attempt to contribute to the biomedical field.

REFERENCES

Alizadeh AA, Eisen MB, Davis RE, Ma C, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 4051: 503-511.

Ben-Abdallah H, Bouassida N, Gargouri, F, Ben-Hamadou, A (2004). A UML-based Framework design method. *J. Object Technol.* 3: 98-119.

Borges HB and Nievola JC (2012). Comparing the dimensionality reduction methods in gene expression databases. *Expert Systems with Applications*. 39: 10780-10795.

Budd TA (2002). An introduction to object-oriented programming. Addison-Wesley, Boston.

Clements P and Northrop L (2002). Software Product Lines: Practices and Patterns. 3 rd edn. Addison-Wesley, Boston.

Froehlich G, Hoover HJ and Sorenson P (2000). Choosing an object-oriented domain Framework. *ACM Computing Surveys*. 32.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.

Hall MA (1999). Correlation-based feature selection for machine learning. Doctoral Thesis, University of Waikato, Hamilton.

Johnson RE (1997). Frameworks = (Components + Patterns). *Communications of the ACM* 40: 39-43.

Johnson RE and Foote B (1988). Designing reusable classes. *Journal of the Object-Oriented Programming* 1: 22-35.

Kent Ridge Bio-Medical Dataset Reprository (2013). Kent Ridge Bio-Medical Dataset. Available at: [http://datam.i2r.a-star.edu.sg/datasets/krbd//]. Accessed March 12, 2013.

Kira K and Rendell LA (1992). The feature selection problem: traditional methods and a new algorithm. Proceedings of Conference on Artificial Intelligence, Menlo Park, 129-136.

Kubo MM and Tori R (2006). FMMG: A Framework for Mobile Multiplayer Games. Proceedings of 8th International Computer Games Conference, Louisville, 64-71.

Liu H and Yu L (2005).Toward integrating feature selection algorithms for classification and clustering. *IEEE Transaction*

on Knowledge and Data Engineering 17: 491-502.
Pohl K, Bockle G and Linden F (2005). Software Product Line Engineering: Foundations, Principles, and Techniques. Springer-Verlag, Berlin.
Prahalad CK and Krishnan M (2008). The New Age of Innovation: Driving Cocreated Value Through Global Networks. 1st edn. McGraw-Hill Professional, New York.
Romdhane LB, Fadhel N and Ayeb B (2010). An efficient approach for building customer profiles from business data. *Expert Systems with Applications* 37: 1573-1585.
Shipp MA, Ross KN, Tamayo P, Weng AP, et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8: 68-74.
Tan CP, Lim KS and Lai WK (2008). Multi-dimensional features reduction of consistency subset evaluator on unsupervised expectation maximization classifier for imaging surveillance application. *Int. J. Image Proc.* 2: 18-26.
Weka (2011). Weka 3: Data Mining Software in Java. Available at: [http://www.cs.waikato.ac.nz/ml/weka/]. Accessed August 25, 2013.