DATA CLASSIFICATION TECHNIQUES FOR CANCER DATASET

Nittaya Kerdprasop, Kittisak Kerdprasop, Prathan Saithong, and Surasiri Noppakan

School of Computer Engineering
Suranaree University of Technology
111 University Ave., Muang District,
Nakorn Ratchasima 30000
Phone (044) 224432, Fax (044) 224165
nittaya@ccs.sut.ac.th, kerdpras@ccs.sut.ac.th, thanlove@msn.com, surasiri@hotmail.com

ABSTRACT

We are flooded with a huge volume of data and information. The tremendous amount of data, collected and stored in large databases, has far exceeded the human ability to analyze and extract valuable information for the purpose of decision-making support. Data mining has emerged as a new technology that can intelligently transform the vast amount of data into useful information and knowledge. Data mining tasks can vary from classification, association, to deviation detection. We focus on the classification technique. The objective of this research is to analyze the different techniques and algorithms of data classification with the intention of discovering the appropriate technique for the cancer dataset. The discovered technique must generate the most accurate classifier with the lowest error rate on predicting the class of unseen data.

1. INTRODUCTION

Enormous amounts of data are being collected daily from scientific projects, stocks trading, hospital information systems, computerized sales records and many other sources. A huge volume of data has far exceeded the human ability to analyze and extract valuable information. This situation has urged for new techniques and automated tools that can intelligently transform the pile of data into useful information and knowledge. Data mining is such an imminent promising technology. The benefit of data mining is to turn the newfound knowledge into actionable results such as increasing a customer's likelihood to buy, or improving the ability to identify patterns of cancer recurrence of patients. We focus on the task of classification, the most extensively studied data mining technique.

Classification is a form of data analysis in that it is the process of extracting models (or patterns) to describe data classes or concepts. The extracted model is used to predict the class of unseen data whose class is unknown. For example, each data item in the dataset gathered from patients who were checked-up for a specific type of cancer was labeled as either negative (no cancer) or positive (having cancer). The extracted model might be the common characteristics and symptoms of most patients who had cancer. This model is useful for the future prediction to determine who is at high risk of having cancer.

Data classification is a two-step process [8]: learning and classification-testing. In the learning phase, data whose classes are known (called the training data) are analyzed by the classification algorithm to build the model. This model can be represented in various forms, for instance, a decision tree, a set of rules, a mathematical formulae. Since the class of

each training data is provided, the classification is categorized as *supervised learning*. In the classification-testing phase, the model is tested on another set of data whose classes are also known (called the test data). The purpose of testing is to estimate the accuracy of the classification model. The process of classification is illustrated in Figure 1.1.

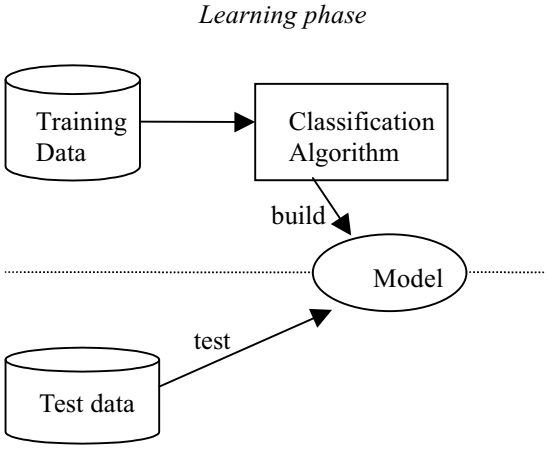

Figure 1.1 The data classification process

Due to the numerous applications of classification (e.g., credit approval, medical diagnosis, stock forecasting), many researchers from diverse disciplines attempt to use various techniques on data classifica-tion. These techniques range from decision tree induction, Bayesian classification, nearest neighbor classification, to case-based reasoning. As the characteristics of data vary from application to application, there is no single technique that performs the best classification on all data types [4, 5, 17, 18].

There is much research in comparing different classification techniques [3, 4, 5, 20]. The team in STATLOG project [18] compares tree-based algorithms against some other classification algorithms on several types of datasets. Another extensive study [17] compares thirty-three classification algorithms. Most comparison studies investigate the algorithms that perform generally well on any kinds of datasets. Our project, on the contrary, emphasizes on the cancer datasets to identify the most appropriate algorithms for this specific domain.

This research compares fourteen different algorithms on two cancer datasets. These datasets are obtained from the UCI Machine Learning Repository [2]. For the purpose of a consistency comparison, we do all experiments in the same environment using the MLC++ system [12]. Each algorithm is compared on the basis of predicting accuracy. The next section explains the datasets used in our experiments. Section 3 briefly describes the classification algorithms. Section 4 outlines the experimental setup. Section 5 reports the results. The last section concludes the paper with some general comments and recommendations.

2. DATASETS

The cancer datasets briefly described in this section are from the UCI Repository [2].

Breast Cancer Dataset

This dataset was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The dataset was reported by M. Zwitter and M. Soklic. The problem is to predict whether a patient who has been treated for breast tumor has recurred-breast-tumor or is safe from the recurrence. The dataset contains 201 instances of no-recurrence class and 85 instances of recurrence class. The instances are described by 9 attributes, four of which are numerical and five are nominal. Nine instances are removed due to the missing values. Our results are thus based on 277 instances.

Lung Cancer Dataset

The data describes three types of pathological lung cancers. The donor is Stefan Aeberhard. He gives no information on the individual variables. The data

contains 32 instances, 56 predictive attributes (all are nominal). The class distribution is 9 instances of class 1, 13 instances of class 2, and 10 instances of class 3.

3. CLASSIFICATION ALGORITHMS

This section describes each classification algorithm briefly. The 14 algorithms are grouped into four categories: basic algorithms (use simple techniques), statistical algorithms, tree-based algorithms, and miscellaneous (use different techniques, e.g., instance-based, decision graph).

3.1 Basic Algorithms

OneR: It is a simple algorithm proposed by Holte [9]. OneR induces classification rules based on the value of a single attribute. OneR is usually used as a base algorithm to compare the predictive accuracy with other sophisticated algorithms. It is shown [9] that we can get reasonable accuracy on many tasks by simply looking at one attribute. The average accuracy of OneR for the datasets tested by Holt is 5.7% lower than that of C4.5.

Const: The algorithm [12] predicts a constant class by simply predicting the majority class in the training data. Although it makes little sense to use this classification scheme for prediction, it can be used as the baseline accuracy to evaluate various classifiers.

Table-majority: A simple table-lookup algorithm [12]. All instances are stored in a table for the purpose of predicting. If an instance is not found, table-majority predicts the majority class of the table.

3.2 Statistical Algorithms

Naïve Bayes: The naïve-Bayes classification algorithm [16] is based on Bayes theorem of posterior probability. Given the instance, the algorithm computes conditional probabilities of the classes and picks the class with the highest posterior. Naïve-Bayes classification assumes that attributes are independence. The probabilities for nominal attributes are estimated by counts, while continuous attributes are estimated by assuming a normal distribution for each attribute and class. Unknown attributes are simply skipped.

Disc-Naïve-Bayes: This is a variant [6] of Naïve Bayes to achieve a better classification by discretizing the continuous attributes. Discretization is performed as a preprocessing step prior to the Naïve-Bayes classification process.

3.3 Tree-Based Algorithms

ID3 and MC4: These are greedy algorithms to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. ID3 [19] is a very basic decision-tree algorithm with no tree-pruning. The algorithm uses an information-theoretic measure to select the attribute tested for each nonleaf node of the tree. MC4 [12] is a decision-tree algorithm with pruning. Pruning is the technique to improve accuracy by removing the branches reflecting noise in the data.

Option decision tree: The tree has option that allow several optional splits, which are then voted as experts during classification [14].

Lazy DT: Lazy decision tree is an algorithm for building the best decision tree regarding each test instance [7].

Nbtree: A decision tree algorithm hybrid with Naïve-Bayes at the level of the leaf nodes [11].

3.4 Miscellaneous

IB: IB is an instance-based (or nearest-neighbor) algorithm [1]. The algorithm stores all training instances and buils the classifier when an unseen instance needs to be classified. The non-trivial computation

is performed in the prediction time to search the pattern closest to the unknown sample.

HOODG: This Hill-climbing Oblivious, read-Once Decision Graph algorithm uses a bottom-up approach to build a decision graph [10] with a hill-climbing technique implemented.

EODG: This is a classification algorithm to build oblivious decision graph top-down [13]. It cannot handle unknown values.

FSS: The Feature Subset Selection is an algorithm that selects a good subset of features (or attributes) for the improved accuracy performance [15].

4. EXPERIMENTS

For each dataset, the experimentation on fourteen classification algorithms has been performed under the same environment, that is, using the MLC++ system [12]. MLC++ is a library of C++ classes and tools supporting supervised learning of concepts. The system provides a variety of tools that help comparing different learning algorithms.

In supervised machine learning, we try to find a set of rules (a classifier) that can be used to accurately predict the class of unseen instance. Thus, the key factor to compare the performance of different classification algorithms is the accuracy. Accuracy estimation is the process of approximating the future performance of a classifier. We use the holdout method to estimate the accuracy. About two thirds of the data are allocated to the training set (for building a classifier), and the remaining (one third) is allocated to the test set. The accuracy on the test set is the estimated accuracy.

5. RESULTS

The classification accuracy of each algorithm on each dataset is reported as the error rate on the test dataset. The results of performance comparison are summarized in Table 5.1 and are also shown graphically in Figure 5.1.

The following conclusions may be drawn from the results:

1. The basic algorithms (i.e., OneR, Const, Table-majority) employ a simple scheme in building a classifier, mostly predicting a majority class. Thus, their performance can be used as a baseline to compare against sophisticated algorithms.
2. The algorithms that perform better (or as good as) the basic algorithms are LazyDT, MC4, OptionDT. These three algorithms are tree-based.
3. The error rates of most algorithms on the lung-cancer dataset are high due to the small size of the dataset.

6. CONCLUSIONS AND DISCUSSION

By the criterion of error-rate comparison, the most accurate algorithms are those in the group of decision-tree induction. The low error rates of IB, EODG, FSS algorithms require further experimentation on a larger dataset. Even though it is natural to measure a classifier's performance in term of the error rate, for the specific domain of medical diagnosis, the cost of missclassification error should be taken into account. Healthy person incorrectly predicted to be ill (false positive) is much less harmful than sick person incorrectly predicted as healthy (false negative).

Therefore, our future plan is to investigate further the decision-tree induction algorithms on a larger dataset with different evaluation methods and comparison criteria.

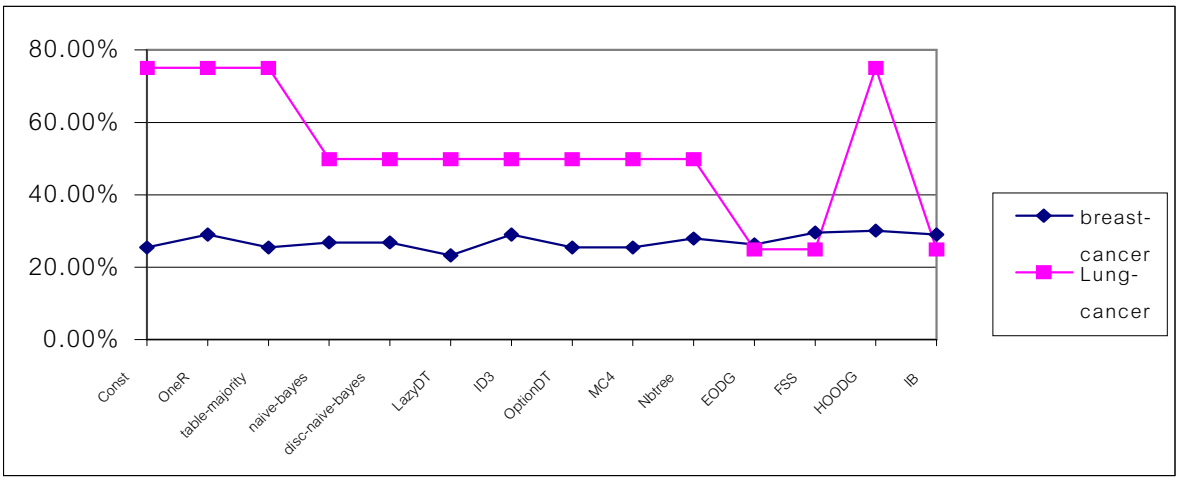

Figure 5.1 Predicting error rate of fourteen algorithms on two datasets

Table 5.1 Error rates of fourteen classification algorithms on two datasets

Dataset	Algorithms						
	Const	OneR	Table-majority	Naïve Bayes	Disc-Naïve-Bayes	LazyDT	ID3
Breast	25.58%	29.07%	25.58%	26.74%	26.74%	23.25%	29.07%
Lung	75%	75%	75%	50%	50%	50%	50%

Dataset	Algorithms						
	OptionDT	MC4	NBtree	EODG	FSS	HOODG	IB
Breast	25.58%	25.58%	27.91%	26.31%	29.47%	30.23%	29.07%
Lung	50%	50%	50%	25%	25%	75%	25%

REFERENCES

[1] D.W.Aha: "Tolerating noisy, irrelevant and novel attributes in instances-based learning algorithms," International Journal of Man-Machine Studies, Vol. 36, No. 1, pp. 267-287, 1992.

[2] C.L. Blake and C.J. Merz: "UCI Repository of Machine Learning Databases," University of California, Irvine, Department of Information and Computer Science, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html]

[3] C.E. Brodley and P.E. Utgoff: "Multivariate versus univariate decision tree," Technical Report 92-8, Department of Computer Science, University of Massachusetts, Amherst, MA, 1992.

[4] D.E. Brown, V. Corruble, and C.L. Pittard: "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problem," Pattern Recognition, Vol. 26, pp.953-961, 1993.

[5] S.P. Curram and J. Mingers: "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," Journal of Operational Research Society, Vol.45, pp.440-450, 1994.

[6] J. Dougherty, R. Kohavi, and M. Sahami: "Supervised and unsupervised discretization of continuous features," Machine Learning: Proceedings of the 12th International Conference, pp.194-202, 1995.

[7] J. Friedman, R. Kohavi, and Y. Yun: "Lazy decision trees," Proceedings of the 13th National Conference on Artificial Intelligence, pp.717-724, 1996.

[8] J.Han and M.Kamber: "Data Mining: Concepts and Techniques,"Morgan Kaufmann, 2001.

[9] R.C.Holte: "Very simple classification rules perform well on most commonly used datasets," Machine Learning, Vol. 11, pp.63-90, 1993.

[10] R. Kohavi: "Bottom-up induction of oblivious, read-once decision graphs," Proceedings of the European Conference on Machine Learning, 1994.

[11] R. Kohavi: "Scaling up the accuracy of Naïve-Bayes classifiers: A decision-tree hybrid," Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp.202-207, 1996.

[12] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger: "MLC++: A Machine Learning Library in C++," Tools with Artificial Intelligence, pp.740-743, 1994.

[13] R. Kohavi and C.-H. Li: "Oblivious decision trees, graphs, and top-down pruning," Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp.1071-1077, 1995.

[14] R. Kohavi and C. Kunz: "Option decision trees with majority votes," Machine Learning: Proceedings of the 14th International Conference, pp.161-169, 1997.

[15] R. Kohavi and D. Sommerfield: "Feature subset selection using wrapper model: Overfitting and dynamic search topology," Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, pp.192-197, 1995.

[16] P. Langley, W. Iba, and K. Thompson: "An analysis of bayesian classifiers," Proceedings of the 10th National Conference on Artificial Intelligence, pp.223-228, 1992.

[17] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih: "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," Machine Learning, Vol. 40, pp.203-229, 2000.

[18] D. Michie, D.J. Spiegelhalter, and C.C.Taylor: "Machine Learning, Neural and Statistical Clasification," Ellis Horwood, 1994.

[19] J.R. Quinlan: "Induction of decision trees," Machine Learning, Vol. 1, pp.81-106, 1986.

[20] J.W. Shavlik, R.J. Mooney, and G.G. Towell: "Symbolic and neural learning algorithms: An empirical comparison," Machine Learning, Vol. 6, pp.111-144, 1991.