A SIMPLE METHOD OF CLASSIFICATION WITH VCL COMPONENTS

Hewen Tang[1], Wei Fang[2] and Yongsheng Cao[2]*

**1: Department of Computer, Yuxi Normal University, Yunnan, China, Email: thw@yxtc.net*

2: Institute of Crop Science, CAAS, Beijing, China, Email: fangweiz@263.net

ABSTRACT

Data mining is a technique to uncover previously unknown and potentially useful knowledge from large datasets. Classification is an important method of data mining. Dew Research has developed some VCL components for classification. In this paper, we implement classification in Delphi with VCL components of Dew Research. It takes well advantage of properties, events and methods of these components and does not need spending much time to consider classification algorithm. With those components, it's also very easy to embody some classification function in some application software dealing with data analysis.

Key words: data mining, classification, VCL component

1 INTRODUCTION

1.1 Data mining

With the progress of humanity society and the development of technique, our capabilities of both generating and collecting data have been increasing rapidly. The widespread use of bar codes for most commercial products, the computerization of many business and government transactions, and the advances in data collection tools have provided us with huge amounts of data. Millions of databases have been used in business management, government administration, scientific and engineering data management, and many other applications. It is noted that the number of such databases keeps growing rapidly because of the availability of powerful and affordable database systems. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically turn the processed data into useful information and knowledge. Consequently, data mining has emerged and become a research area with increasing importance[1].

Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data[1]. It has many alias, such as knowledge discovery in database(KDD), pattern discovery, Knowledge extracting, data dredging, data archeology and so on[2].

Data mining is usually used to but not limited to business[3]. Both major parties in the 2004 U.S. election utilized data mining of potential voters[4]. Data mining has been heavily used in the medical field, to include diagnosis of patient records to help identify best practices[5]. The Mayo Clinic worked with IBM to develop an online computer system to identify how that last 100 Mayo patients with the same age, gender and medical history had responded to particular treatments[6].

Data mining has a lot of methods such as classification, cluster and association rule. This paper focuses on classification.

1.2 Classification

Classification is one of the most common data mining techniques, which is used to predict group membership for data instances. In order to understand and communicate about the world, we are constantly classifying, categorizing,

and grading. For example, we divide living things into phyla, species, and general; matter into elements; dogs into breeds; people into races.

A class is a set of data samples with some similarity or relationship and all samples in this class are assigned the same class label to distinguish them from samples in other classes[7].

Classification consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. The objects to be classified are generally represented by records in a database table or a file, and the act of classification consists of adding a new column with a class code of some kind. The classification task is characterized by a well-defined definition of the classes, and a training set consisting of preclassified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it.

Classification has a wide spread application in actual world. For example, it can predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy". Popular classification techniques include decision trees and neural networks.

1.3 Delphi

Delphi is an object-oriented, visual programming environment to develop applications for deployment on Windows and Linux[8]. Usually, Delphi is considered as Rapid Application Development(RAD) tools. It owns a very large visual component library(VCL), which provides large amount of various components. This is very helpful for developing highly efficient applications with a minimum of manual coding.

Besides its own components, Delphi also allows to add some components form third part to the component palette. This expands largely its own function and make developing application more easily and efficient. This paper is a good example.

1.4 Dew Research's VCL Components of data mining

Dew Research is a software company founded in 1999 in Slovenia. It develops numerical, statistical and digital signal processing libraries, which scientists and engineers need to build standalone royalty free Microsoft Windows applications. It focuses on the development of reusable parts of applications, components and patterns[9].VCL Components of data mining is one of the main products of Dew Research. It consists of four components, which are NaiveBayes, K-Nearest Neighbours, Linear Classifier and DBClassifier.

These components are not alone and it's included in Dew Lab Studio Demos which can be downloaded from www.dewresearch.com. It is not free but its trial edition is available free for every one from the website of Dew Research.

The first step to use these VCL Components is to install it on computer. Once the Dew Lab Studio Demos installed, these components are embedded in the component palette of Delphi as components of page DewAI as Fig.1.

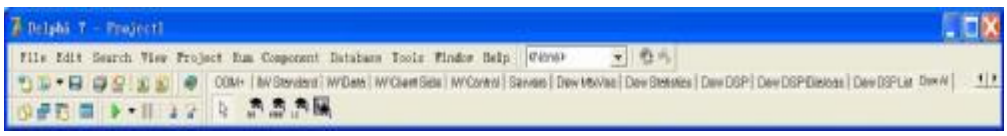

Figure 1. VCL Components of data mining embedded in Delphi

2 DATA PREPARATION

We select dataset of Large Soybean Database[10].

This dataset has 35 categorical attributes---- date, plant-stand, precip, temp, hail, crop-hist, area-damaged, severity, seed-tmt, germination, plant-growth, leaves, leafspots-halo, leafspots-marg, leafspot-size, leaf-shread, leaf-malf, leaf-mild, stem, lodging, stem-cankers, canker-lesion, fruiting-bodies, external decay, mycelium, int-discolor, sclerotia, fruit-pods, fruit spots, seed, mold-growth, seed-discolor, seed-size, shriveling, roots.

These attributes are some nominal and some ordered, for example the attribute of "hail" is nominal whose values are "yes" or "no", and the attribute of "germination" is ordered whose values are 90-100%, 80-89%, lt-80%. This can not fit the classification algorithm directly. This paper have values for attributes encoded numerically, with the first value encoded as ``1, '' the second as ``2,'' and so forth. For example, the values of "hail" are numerically coded as "1" or "2". The final dataset is show as Fig2.

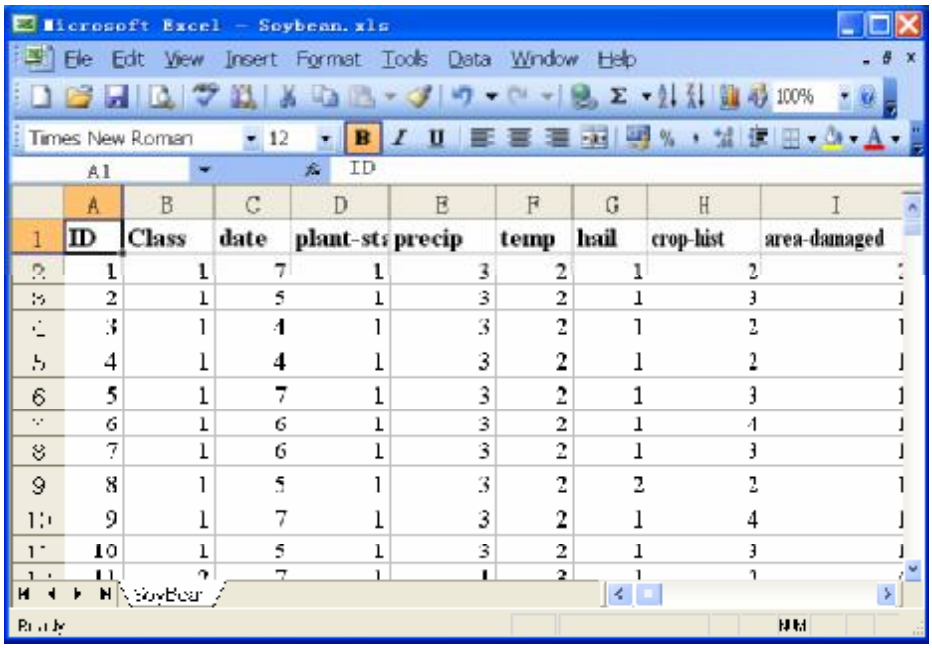

Figure 2. Dataset of Large Soybean Database

There is a new attribute of "Class", which is the actual class and only used to testify the result of classification.

The dataset certainly could be prepared as a table in database. This paper prepared dataset in database. It is shown here as Excel file only to offer an example of single file except database.

3 CLASSIFICATION WITH VCL COMPONENTS OF DEW RESEARCH

We implemented a data mining task, classification, in Delphi with VCL Components of Dew Research. This approach need only few code to implement classification without spending much time in algorithm except calling some methods or events of VCL Components.

3.1 Arrange components in a form

First step of this approach is to place some components on a form of Delphi. Here all four of data mining VCL Components of Dew Research are placed on a form. Besides this, two database components (DataSouce, Table) and some other usual components are also placed on the same form. The final arranged form is shown as Fig3.

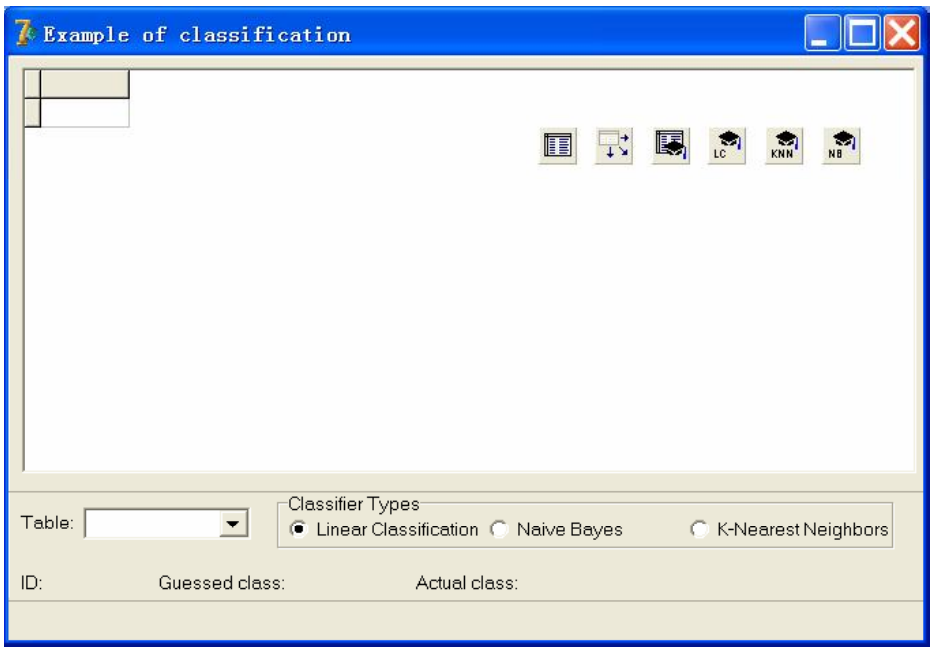

Figure 3. Arrangement of components

3.2 Implement of classification

Classification is often divided into two steps----learning and classifying. This paper calls some methods of data mining VCL Components of Dew Research to implement two of them. Some important methods are listed in Tab1.

Table 1. Some mainly important methods and their function

Name	Function
LearnData	Teach the classifier attached to the Classifier property with the data in the dataset.
Learn	Teach the classification class
Classify	Determine the most probable class to which the case belongs
ClassifyRecord	Classify the current record and return the name of the class.

3.3 Result of classification

The result of implemented classification in this paper is show as Fig4.

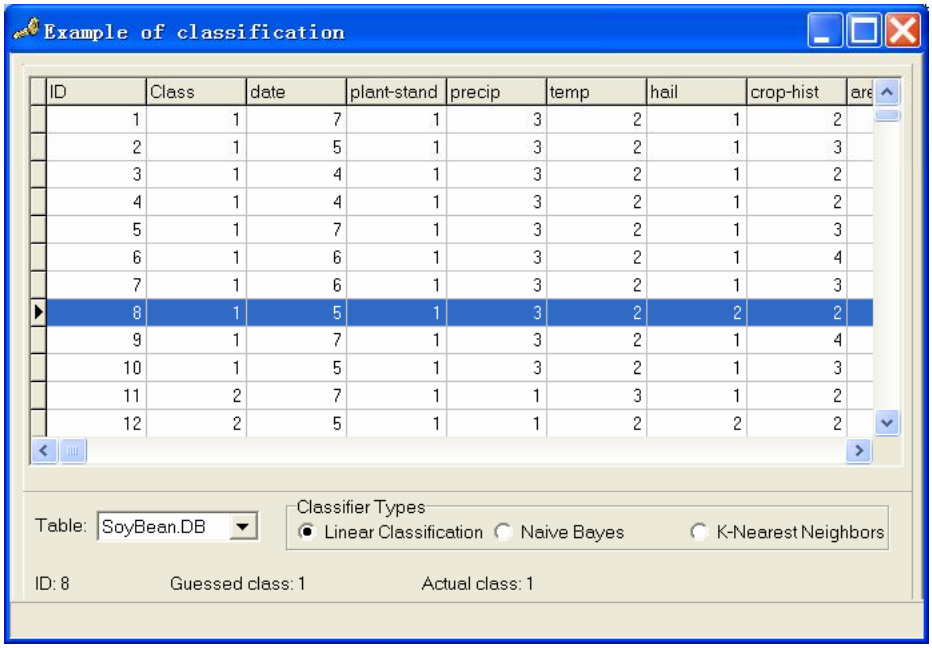

Figure 4. Arrangement of components

In this result, once a record is clicked, its correspondent classification result is given out in the bottom of Fig4.

4 RESULT AND DISCUSSION

The method in this paper shows that it's simple to classify with these data mining VCL Components of Dew Research. User does not need much time to implement classification algorithm but only need to call some method.

This paper presents only the main work of classification with these data mining VCL Components. Much work of data preparation is omitted, such as establishing database, filling data to components. Additionally, these components can also work with data file besides database.

The method of this paper is only one simple possible implement of classification with these data mining VCL Components. For a well complete application of classification, there is much to do, such as improving user interface, enriching the dataset selection.

5 REFERENCE

[1] U.M.Fayyad, G.Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy(1996). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

[2] H.David, M.Heikki and S.Smyth(2001). Principles of Data Mining. The MIT Press

[3] H.Jiawei, K.Micheline(2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc

[4] L.O.David, D.Dursun(2008). Advanced Data Mining Techniques, Springer-Verlag Berlin Heidelberg.

[5] H.Havenstein(2006). IT efforts to help determine election successes, failures: Dems deploy data tools; GOP expands microtargeting use, Computerworld 40:45, 1, 16.

[6] T.G. Roche (2006). Expect increased adoption rates of certain types of EHRs, EMRs, Managed Healthcare Executive 16:4, 58.

[7] Lipo Wang, Xiuju Fu(2005). Data Mining with Computational Intelligence. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

[8] Borland Software Corporation (2001). Delphi 7 Developers' Guide. Borland Software Corporation

[9] www.dewresearch.com

[10] R.S. Michalski and R.L. Chilausky(1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. International Journal of Policy Analysis and Information Systems, 4:125-161.