AUGMENTATION MODERNISM APPLICATIONS IN AGRICULTURE USING DATA MINING

Dr.V.MURUGAN
Head , Dept of Computer Science
Shanmuga Insutries Arts & Science College
Tamil Nadu , India.
profmuruga@gmail.com

Abstract

The system provides a comprehensive suite of facilities for applying data mining techniques to large data sets. This paper discusses a process model for analyzing data, and describes the support that system provides for this model. The domain model 'learned' by the data mining algorithm can then is readily incorporated into a software application. This system based analysis and application construction process is illustrated through a case study in the agricultural domain—mushroom grading.

Keywords: machine learning, data mining, data analysis, application development

INTRODUCTION

Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets (Piatetsky-Shapiro and Frawley, 1991). The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. Alternatively, human domain experts may choose to manually examine the model, in search of portions that explain previously misunderstood or unknown characteristics of the domain under study. In our work, we concentrate on machine learning techniques for inducing domain models or analyzing datasets (described further in Section 3). Machine learning algorithms provide models with a classification/prediction accuracy comparable to, for example, artificial neural networks, but which are more intelligible to humans than a neural model. The WEKA1 research team has two objectives: to mine information from existing agricultural datasets produced by New Zealand scientists and research organizations; and to perform basic research in data mining by developing new machine learning algorithms. To support these goals, we have developed a data mining workbench, the WEKA system, that incorporates the following tools: a set of *data pre-processing routines*, supporting the manipulation of raw data and its transformation into an appropriate form for

data mining; *feature selection tools*, useful for identifying irrelevant attributes to exclude from the dataset; classifiers and other *data mining algorithms*, capable of handling categorical and numeric learning tasks; *metaclassifiers* for enhancing the performance of classification data mining algorithms (for example, boosting and bagging routines); *experimental support* for verifying the comparative robustness of multiple induction models (for example, routines measuring classification accuracy, entropy, root-squared mean error, cost-sensitive classification, etc.); and *benchmarking tools*, for comparing the relative performance of different learning algorithms over several datasets.

DATA MINING PROCESS MODEL

In the course of this project we have analyzed over 50 real-world data sets, primarily agricultural data sets provided by research institutes and businesses in New Zealand. From this experience we have developed a process model for applying data mining techniques to data, with the goal of incorporating the induced domain information into a software module (Figure 1). The key points of this model are (Garner et al, 1995): • *a two-way interaction between the provider of the data and the data mining expert*. Both work together to transform the raw data into the final data set(s) input to the machine learning algorithms — with the domain expert providing information about data semantics and 'legal' transformations that can be applied to the data, and the data mining expert guiding the process so as to improve the intelligibility and accuracy of the results.

• *an iterative approach.* Machine learning is an exploratory process; it generally takes several cycles through the process model to find a good "fit" between a representation of the data and a data mining algorithm. In addition, distinct attribute combinations run through different schemes can produce wildly different data models, even though the predictive accuracy of the results may be equivalent. These alternative views may provide valuable insights into patterns covering different subsets of the data. In the model presented in Figure 1, activity flows in a clockwise direction. In the preprocessing stage, the raw data is firstly represented as a single table, as required by the data mining algorithms included in WEKA. This table is translated into the ARFF format, an attribute/value table representation that includes header information on the attributes' data types. The data may also require considerable 'cleansing', to remove outliers, handle

missing values, detect erroneous values, and so forth. At this point the data provider (domain expert) and the data mining expert collaborate to transform the cleansed data into a form that will produce a readable, accurate data model when processed by a data mining algorithm. These two analysts may, for example, hypothesize that one or more attributes are irrelevant, and set aside these extraneous columns. Attributes may be manipulated mathematically, for example to convert all columns containing temperature measurements to a common scale, to normalize values in a given column, or to combine two or more columns into a single derived attribute.

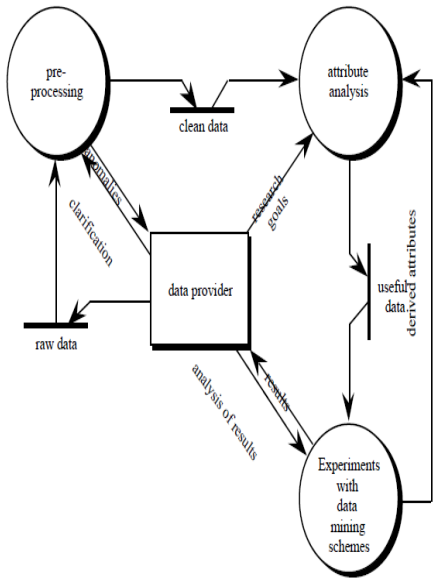

Figure 1. Process model for a machine learning application (data flow diagram)

One or more versions of the cleansed data are then processed by the data mining schemes. The domain expert determines which portions of the output are sufficiently novel or interesting to warrant further exploration, and which portions represent common knowledge for that field. The data mining expert interprets the algorithms' output and gives advice on further experiments that could be run with this data.

MACHINE LEARNING TECHNIQUES

The current version of WEKA contains implementations of twelve learning schemes: ten classifiers, a clustering algorithm, and an association rule learner. The software architecture is flexible enough to permit other learning schemes, and other types of learning schemes, to also be slotted into WEKA. In this section, we describe the types of learning that WEKA currently supports.

Classifiers

The output from this type of learning scheme is, literally, a classifier—usually in the form of a decision tree or set of rules that can be used to predict the classification of a new data instance. One attribute in the input table is designated as the category or class for prediction; the rest of the attributes may appear in the "if" portions of the rules (or the nonleaf nodes of the decision tree).

Meta-Classifiers

Recent developments in computational learning theory have led to methods that enhance the performance or extend the capabilities of these basic learning schemes. We call these performance enhancers "meta-learning schemes" or "meta-classifiers" because they operate on the output of other learners. Instead of using a single classifier to make predictions, why not arrange a committee of classifiers to vote on the classification an instance? This is the basic idea behind combining multiple models to form an ensemble or meta classifier.

Two of the most prominent methods for constructing ensemble classifiers are boosting and bagging (Breiman, 1992). More often than not, these classifiers can increase predictive performance over a single classifier. However, the price for this increase in performance is that it is generally not possible to understand what is behind the improved decision making.

Clustering

Clustering methods do not generate predictive rules for a particular class, but rather try to find the natural groupings (or "clusters") in the dataset. This technique is most often used in an exploratory fashion, to generate hypotheses about the relationships between data instances. Clustering is often followed by a second learning stage, in which a classifier is used to induce a rule set or decision tree that allocates each instance in the dataset to the cluster assigned to it by the clustering algorithm. These classifier-generated 'cluster descriptions' can then be examined to gain a semantic understanding of the clusters. WEKA includes an implementation of the EM clustering algorithm. This algorithm makes the assumption, common to other clustering algorithms, that the attributes in the dataset represent independent random variables. Some clustering algorithms force each record to belong to exactly one cluster; EM permits an instance to belong to more than one cluster, a useful extension that, in practice, can support more flexible and more 'fuzzy' descriptions of the implicit structure of the dataset.

WEKA TOOLS

In addition to the learning algorithms discussed above, WEKA also provides tools for preprocessing data and for comparing the performance of different learning algorithms.

Dataset pre-processing

WEKA's pre-processing capability is encapsulated in an extensive set of routines, called *filters* that enable data to be processed at the instance and attribute value levels.

These filters have a standard command-line interface with a set of common command-line options.

Many of the filter algorithms provide facilities for general manipulation of attributes—for example, to insert and delete attributes from the dataset. When experimenting with learning schemes in the development of a data mining application (Section 2), one of the most common activities involves building models with different subsets of the complete attribute set. WEKA provides three feature selection systems to aid in choosing attributes for inclusion in an experiment: a locally produced correlation based technique (Hall and Smith, 1998); the wrapper method (John and Kohavi, 1997); and Relief (Kira and Rendell, 1992).

CONCLUSION

As illustrated by the case study presented in Section 5, information 'mined' from data can provide insights into the domain being studied that run counter to the received wisdom of a field. Locating these surprising or unusual portions of the model can be the focus for a data mining analysis, so that the results can be applied back in the domain from which the data was drawn. In this case, the results indicate that the subjective attributes for mushroom grading may not be useful in practice, and so perhaps they need not be measured or recorded. Criteria based on the attributes found in the J4.8 models may prove useful in developing more objective standards for quality classification and market pricing for mushrooms.

In other data mining applications, the goal might be to use a model predictively, to provide automated classification of new instances. In these applications, the learning component will likely be a small part of a much larger software system. Since WEKA learning schemes are accessible from other programs, a learning module can be slotted into a larger system with a minimum of additional programming.

REFERENCES

1. Agrawal, R., Imielinski, T., and Swami, A. (1993) "Mining association rules between sets of items in large databases." *Proceedings of the ACM SIGMOD Conference onManagement of Data*, Washington, D.C., 207-216.

2. Agrawal, R., Imielinski, T. and Swami, A.N. (1993) "Database mining: a performance perspective." *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5,914–925.

3. Aha, D. (1992) "Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms." *International*

Journal of Man-Machine Studies, Vol. 36, 267–287.

4. Atkeson, C.G., Schaal, S.A. and Moore, A.W. (1997) "Locally weighted learning." *AIReview*, Vol. 11, 11–71.

5. Blake, C.L. and Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

6.Fayyad, U.M. and Irani, K.B. (1993) "Multi-interval discretization of continuous-valued attributes for classification learning." *Proceedings of IJCAI*, Chambery, France, 1022–1027.

7. Friedman, J.H., Hastie, T. and Tibshirani, R. (1998) "Additive logistic regression: a statistical view of boosting." Technical Report, Department of Statistics, Stanford University.

8.Freund, Y. and Schapire, R.E. (1996) "Experiments with a new boosting algorithm."

Proceedings of COLT, 209–217. ACM Press, New York.