Exploratory Attributes Search for Time-Series Data: An Experimental System for Agricultural Application

Kazunori MATSUMOTO

System Integration Technology Center,
TOSHIBA,
3-22 Katamachi, Fuchu, Tokyo, 183-8512, Japan
E-mail: kazunori@sitc.toshiba.co.jp

Abstract. This paper reports an experimental agricultural datamining system which purposes to find weather patterns influencing yield of rice. Necessary data for this system are separately maintained in various databases. We then show how this system integrate them into one database with an assistance of support databases. Next we discuss the attribute selection problem for the data in the integrated database. Our method first exploratory search for a candidate set of attributes. In this case, the support databases is used to avoid a searching space explosion. Once the candidate set is identified, we apply a greedy search in the set to find the most useful subset of attributes.

1 Introduction

Agriculture is an information-intensive industry from an essential point of view. So many factors such as soil, fertilizer, temperature, precipitation, sunray, etc. are all affect harvest, so that information about them is carefully investigated by expert persons in deciding agricultural activities. We thus expect to build up an intelligent computerized agricultural information system [7] to assist the experts and to help an improvement on agricultural technologies. Towards this purpose, we firstly need to provide a system which can reveal hidden relations among agricultural factors. Although traditional statistical methods have already applied to this field, we expect recent datamining technologies [1] to bring still more fruitful results. For example, an expert can easily understand IF - THEN style rules extracted by the typical datamining methods, then he may give further investigation around the rules. In this paper we reports an experimental agricultural datamining system whose purpose is to find weather patterns determining yield of rice. Necessary data for this datamining are maintained in separated databases. We then need to integrate them into one database. Since each database is built independently, their integration cannot achieve in direct way. We show how they are integrated by using support databases which store additional agricultural information.

Most datamining methods [1,6] run with a set of training data which is specified as a set of tuples of attribute values with class information. Then we must identify the set of attributes, and transform data in the integrated database with the attributes. This attribute selection problem is actively studied [3,4,5], however, previous works mostly concern the method of removing unimportant attributes from the initially given ones. We show the usual attribute selection is not adequate for our purpose.

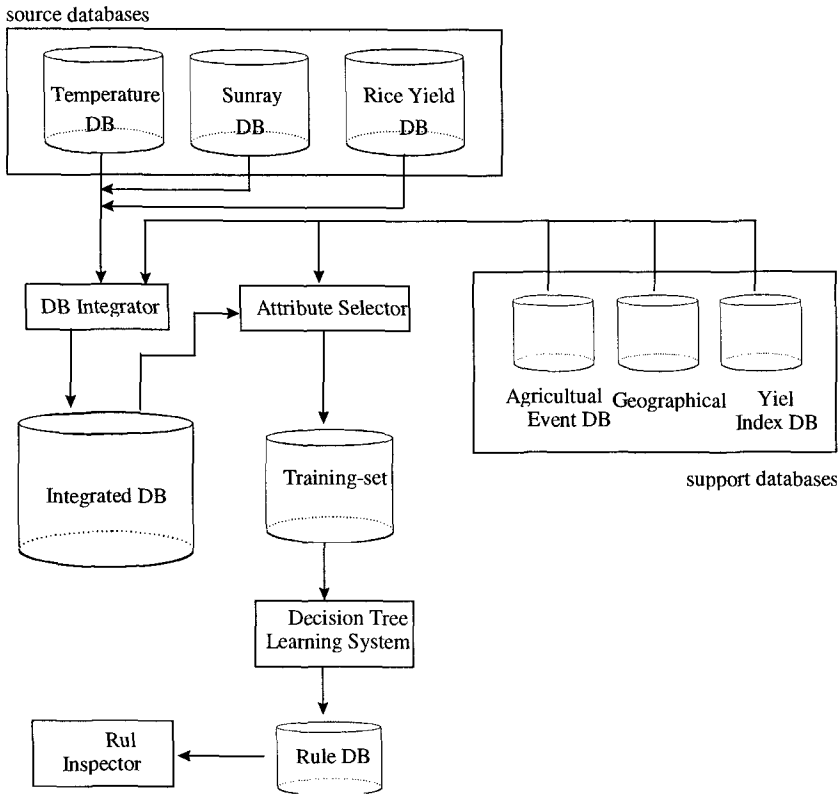

Fig. 1 Outline of the experimental system

2 Outline of the system

According to the general flow shown in Fig.1, the system arranges databases and analyze them by using datamining techniques. The raw data necessary for the datamining phase are separately maintained in the left above three source databases.

The DB Integrator makes the Integrated DB which stores all necessary information in an unified form. In this case, the DB Integrator uses the support databases to bridge the gap among source databases. The Attribute Selector searches in the Integrated DB for an adequate set of attributes with an assistance of the support databases. Data in the Integrated DB are re-expressed in terms of the attributes and are stored in the Training-set DB. The Decision tree learning system datamines the training-set DB and extracted classification rules which are stored in the rule DB. The Inspector with the Rule DB re-analyze the training-set DB and the Integrated DB to proceed further examination of the rules.

3 Making the Training-set database

As we seen in the outline, the training-set DB is built via the Integrated DB. In this chapter, we explain how the Training-set database is made with the support databases.

3.1 Integrating the source databases

In this experiment, we use the following three source databases. Before integration, they are separately arranged by using the support databases.

Rice Yield DB
This database records, in tabular form, yearly yield of rice by the kilogram per ten are (kg/10a) for each prefecture in Japan. The schema is given as:
<Prefecture Name, Year, Yield>. Fig.2 graphically shows a part, Hokkaido which is the northmost prefecture in Japan, of this databases. From this graph, we see the yield tend to increase in general with sudden oscillation. The general increasing tendency is mainly brought by the progress of agricultural technologies in the long period. On the other hand, sudden rises and falls mostly due to the natural conditions, especially weather is one of the most affective factors. For the purpose of our experiment, we need to remove the contribution of the technological improvement from the graph. In case of rice, the Japanese Agricultural Ministry publishes the yield index every year, which are stored in the Yield Index DB. Before the integration, we rewrite the Yield value in theYield DB with the index.

Temperature DB and Sunray DB
We use these two weather databases in this experiment. These values are observed at several weather stations in each prefecture, and their daily averages are recorded in each databases. The both databases are in tabular form and their schema is common as: <Weather Station Name, Year, Value of Jan.1, ..., Value of Dec.31>.

The join operation [8] provides a standard way of integrating several databases in tabular forms into a new tabular form database. This operation combines two tuples when they match on the join columns. In case of our experiment, we need to resolve the difference of Prefecture Name and Weather Station Name before applying the join operation. The Geographical DB is used for this purpose.

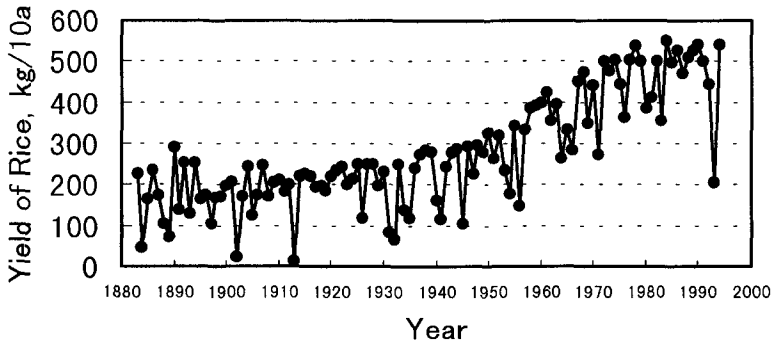

Fig. 2 Yearly yield at Hokkaido Prefecture

Geographical DB(support database)
This database stores the geographical information about the weather stations. The data schema is: <Weather Stations Name, Prefecture Name, Altitude,Longitude, Latitude>. With this information, we can get all weather station names for a given prefecture. Then, temperature (sunray) data for a prefecture X can be calculated as the average of the values at all weather stations in X (Fig.3). In this case, a station of which altitude considerably different from others is ignored. Outliers can also be removed in this step by ignoring extreme values. Similarly, a data lack at a station can be compensate with other station's value.

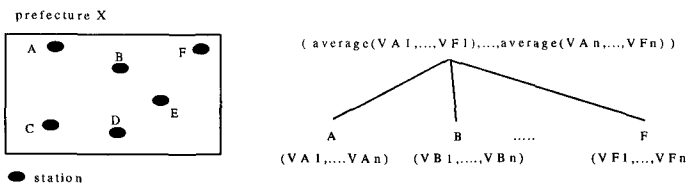

Fig. 3 Prefecture Data is Calculated from Station Data

Finally, we can integrate above three source databases into the integrated DB, whose schema is: <Prefecture Name, Year, Arranged Yield, Temperature of Jan.1, ..., Temperature of Dec.31, Sunray of Jan.1, ..., Sunray of Dec.31>.

3.3 Exploratory Search for Attributes

Most datamining tools assume a set of training data is specified as a set of tuples of attribute values with class information. A typical datamining is a process to build good rules in terms of attributes or attribute values, which correctly classifies most of the training data. It is reported [3,4,5] that inadequately selected attributes cause a searching space explosion and cause decreasing the extracted rule quality. Then, defining a set of suitable attributes is an important issue. Most studies concerning this issue formalize it as the attribute subset selection problem, which identifies unrelevant (or unimportant) attributes from given initial set of attributes.

A direct application of the subset selection approach to our case regards every time points as the initial attributes, then remove unrelevant time points from the initial set. This direct method is not enough for our case because of the following reasons:

(1) This is unaware of a data behavior on an time interval. In case of rice growing, existing agricultural knowledge says that weather patterns on a relatively long interval are important rather than momentary patterns.

(2) Date of agricultural activities or events, such as seeding, harvesting, are changed year by year (Fig.4). A direct comparison of different years has few agricultural meaning.

Then, we successively enumerate a time intervals, as a candidate for an attribute, by focusing on the meaningful date specified in the Agricultural Event DB, then check its importance in exploratory manner. In this case, we introduce heuristics to avoid generating too many intervals. First, we can immediately exclude portions which are obviously unrelated to rice growing. It starts at the day of seeding (SD, for short), usually at the middle of April, and ends at the day of harvesting (HD, for short), usually at the end of September, then it is enough to focus on this interval [SD,HD]. Second, since all of SD,TD,ED,and HD are used as milestones in planning agricultural activities, we restrict at least one boundary of an interval must be one of these date. Third, from agricultural commonsense, an interval length can be set greater than two weeks.

According to above the heuristics and the following criterion, we search for a set of useful intervals.

Criterion: We say an interval is useful, if the class difference of the training data can be well explained by their difference on the interval.

In this experiment, the class difference of two training data is defined as the difference of their class values, and the difference of two training data is defined as the difference of their average value on the interval. According to these definitions, we exploratory investigate the data for searching useful intervals. Currently, the main work of this stage is various kinds of visual inspection of data with an assistance of statistical information.

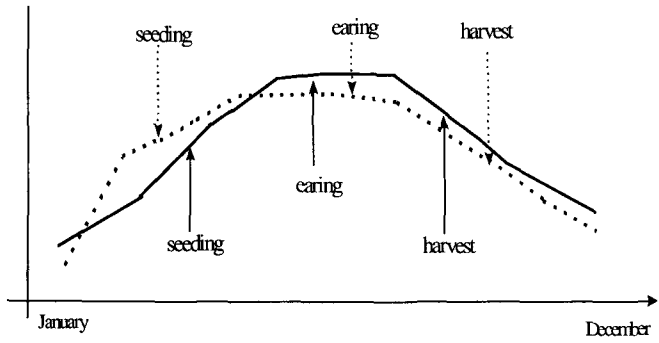

Fig. 4 Agricultural Events Differ Year by Year

The left side of Fig. 5 shows an example of visual investigation on the interval [SD,TD]. The X-axis is the training data difference and the Y-axis is the class difference. The right side of the figure shows the similar investigation on the interval [ED, ED+3weeks]. In this case, we conclude [ED,ED+3weeks] is useful, on the other hand [SD,TD] is regarded as not useful. Continuing this investigation, we finally select 8 intervals (4 is for temperature data, and others are for sunray data)as the attributes set.

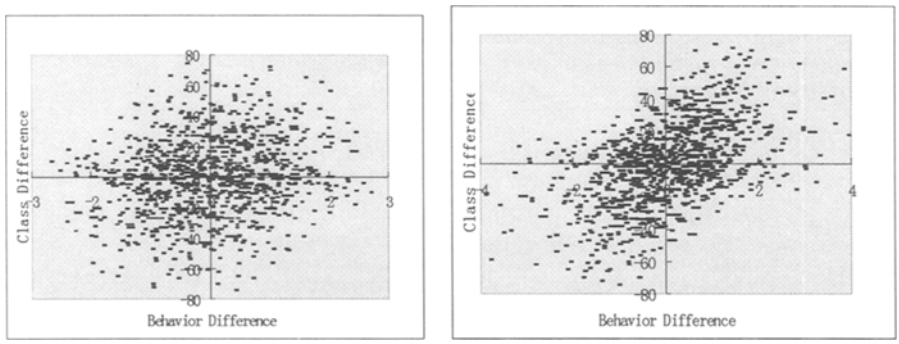

Fig. 5 Class difference and data difference on intervals

4 Mining the training-set database

Once the training-set database has been prepared, we datamine it with the decision tree learning system. Currently we use the C4.5 [6], which is one of the most successful datamining tool. Due to the limitation of the C4.5, we replace the continuous class values by discrete values of 'good' and 'bad'. By an direct application with the training-set DB, we can extract a set of rules, which have 20.9% of estimated error rate. This is not satisfactory result. Here each element of the current attribute set is checked its usefulness by the visual exploration, however, usefulness of their combination is not yet verified. We then use greedy attribute selection over the current set of attributes. That is, we repeatedly run the C4.5 with every subset of the current attribute set, and select the best performed one. Since the exploratory selection stage focuses sufficiently small attributes, computational cost of the greedy selection can be ignored. Fig. 6 shows the results of the greedy selection. The X-axis shows subset rank in order of goodness. The Y-axis is the estimated error rate with the subset. We finally get a set of rules with 14.2 % estimated error rate, which is 6.2% improvement over the first application.

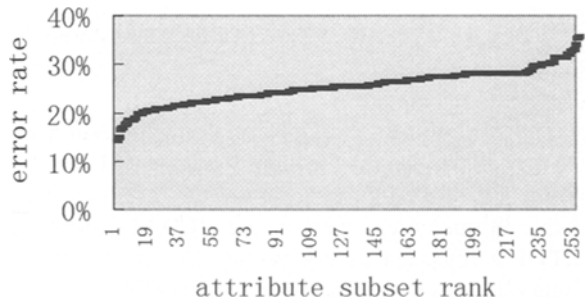

Fig.6 error rate for each attribute subset

5 Conclusion

We report an experimental agricultural datamining system. From a database perspective, method for integration of all necessary data from separately maintained various databases is an important issue. Recent studies on dataware house [2] provide applicable technologies, however, further improvement is inevitable for our purpose.

We plan to combine other datamining technologies into this experimental system to get more fruitful results.

Acknowledgment
The author would like to thank to Dr. Hiroshi SEINO of National Institute of Agro-Environmental Sciences for providing essential data and giving many suggestive advice.

References

[1] M.S. Chen, J. Han, and P.S. Yu, Data Mining: An Overview from a Database Perspective, IEEE Trans. on Knowledge and Data Engineering, Vol.8, No.6, 1996.
[2] R. Kimball, The Data Warehouse Toolkit, John wiley & Sons,1996
[3] K. Kira, and L.A. Rendell, The Feature Selection Problem: Traditional Methods and a New Algorithm, *Proc. of The Ninth National Conf. on AI*, 1992.
[4] R.Kohavi and D.Sommerfield, Feature Subset Selection using the Wrapper Model:
Overfitting and Dynamic Search Space Topology, First Int. Conf. on KDD, 1995.
[5] H. Liu, and R. Setiono, A Probabilistic Approach to Feature Selection - A Filter Solution , *Proc. of The Thirteenth Int. Conf. on ML*, 1996.
[6] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[7] R.B. Rao et.al, Data Mining of Subjective Agricultural Data, Proc. of the Tenth Int. Conf. Machine Learning, 1993.
[8] J.F. Ullman, Principles of Database Systems, Computer Science Press, 1982.