Use of Data Mining Tools in the Fields of Tea Cultivation and Tea Industry of Assam

Sadiq Hussain
System Administrator
Examination Branch
Dibrugarh University

Nayeemuddin Ahmed
Assistant Professor
Centre for Computer Studies
Dibrugarh University

ABSTRACT
Data mining has great potential in the fields of tea cultivation and tea industry of Assam for exploring the hidden patterns in the data sets of the domain. These patterns can be utilized for tea cultivation analysis. However, the available raw data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form an information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only,where data mining deals with heterogeneous fields.

Keywords
Data mining, Apriori Algorithm, Association Rule

1. INTRODUCTION
With the tremendous improvement in the speed of computer and the decreasing cost of data storage, huge volumes of data are created. However, data itself has no value. Only if data can be changed to information, it becomes useful. In order to generate meaningful information, or knowledge from database, the field of data mining was born. The data mining field is about two decade old. Early pioneers such as U. Fayyad, H. Mannila, G. Piatetsky-Shapiro, G. Djorgovski, W. Frawley, P. Smith, and others found that the traditional statistical techniques[9] were not adequate to handle the mass amount of data. They recognized the need of better, faster and cheaper ways to deal with the dramatic increase in the amount of data. Nowadays, besides the numerous number of databases created and accumulated in a dramatic speed, data is no longer restricted to numeric or character only especially in the industry and cultivation aspect. The advanced cultivation techniques, devices and database management systems enable the integration of the different types of high dimensional multimedia data (e.g. text, image, audio, and video) under the same umbrella. Establishing a methodology for knowledge discovery and management of large amounts of heterogeneous data has therefore become a main priority. In this paper we present our investigation results of the applications of the data mining in the tea cultivation aspect, which includes the area of biology, industry and economy.

2. MATERIALS AND METHODS
In this paper, we are creating tables containing information about all the years (breaking down into months) with the fields - rainfall, humidity, temperature and calculating amount produced by using association rule[6] and then apply the Apriori algorithm to see how the factors have effect on the production. The procedure of having the data marked as low,medium, high.

For making a field survey on Tea plants and Tea industry, we had visited a number of identified district of Assam in various tea gardens of Assam and Research organization such as PanitolaTea Estate of Tinsukia District,Chabua Tea Estate of Dibrugargh District ,Moran Tea Estate of Sibsagar District, Bargang Tea Estate of Sonitpur,Monabari T.E of Sonitpurand Dikom Tea Research Center.We personally interact with all the managers and co-workers of various tea gardens.I also personally interact with head of the Dikom Tea Research association(TRA) to collect data and important information required in accomplishing my task. In addition, we had inputs from local reports, communications and scientific publications from different areas of north-east India.

3. RESULTS AND DISCUSSION
In computer science and data mining, Apriori[3] is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

As is common in association rule mining[8], given a set of itemsets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Table I. Production Table based on Data Collected

Month/ Year	**2005**	**2006**	**2007**	**2008**	**2009**
Jan	0	0	0	0	0
Feb	0	1599	1342	1185	3636
March	151010	189124	262660	133154	36596
April	353315	187321	273449	459444	350557
May	400543	494598	418597	454582	413541
June	541984	492081	605315	633638	614299
July	848446	799747	848820	850948	775275
Aug	936297	806514	816914	890857	961652
Sep	770740	882183	937639	824813	876388
Oct	861222	812251	832204	931347	925059
Nov	463008	424444	475977	459067	524083
Dec	127898	116789	106581	80184	122738

January Month

LT=low temperature LT range=9°--25°C

LR=low rainfall LR range=0—10cm

LH=low humidity LH range= morning(83-87%) and afternoon(85-88%)

Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	LT,LR,LH
2006	LT,LR,LH
2007	LT,LR,LH
2008	LT,LR,LH
2009	LT,LR,LH

1-Itemsets	Sup-count
LT	5
LR	5
LH	5

Prune

C1

1-Itemsets	Sup-count
LT	5
LR	5
LH	5

Join

2-Itemsets	Sup-count
LT, LR	5
LT, LH	5
LR, LH	5

Prune

Frequent 2-Itemsets	Sup-Count
LT, LR	5
LT, LH	5
LR, LH	5

Join

Frequent 3-Itemsets	Sup-count
LT, LR, LH	5

Applying Association Rule:

Temperature (X,'9---25') ^ Rainfall (X,'0---10') ^ Humidity (X,'83---87' & X,'85---88)
Production (X,'0')

February Month

LT=low temperature LT range=9°--25°C
LR=low rainfall LR range=0—10cm
LH=low humidity LH range= morning(83-87%) and afternoon(85-88%)
MT=medium temperature MT range=13°--29°C

Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	MT,LR,LH
2006	MT,LR,LH
2007	MT,LR,LH
2008	MT,LR,LH
2009	MT,LR,LH

1-Itemsets	Sup-count
MT	5
LR	5
LH	5

Prune

C1

1-Itemsets	Sup-count
MT	5
LR	5
LH	5

Join

Frequent 2-Itemsets	Sup-count
MT,LR	5
MT,LH	5
LR, LH	5

Prune

Frequent 2-Itemsets	Sup-Count
MT, LR	5
MT, LH	5
LR, LH	5

Join

Frequent 3-Itemsets	Sup-count
MT, LR, LH	5

Applying Association Rule:

Temperature (X,'13---29') ^ Rainfall (X,'0---10') ^ Humidity (X,'83---87' & X,'85---88') ⟹
Production (X,'1k---4k') where k=1000

March Month

LT=low temperature LT range=9°--25°C
MT=medium temperature MT range=13°--29°C
LR=low rainfall LR range=0—10cm
MR=medium rainfall MR range=10---20
LH=low humidity LH range= morning(83-87%) and afternoon(85-88%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	MT,MR,LH
2006	MT,LR,LH
2007	MT,LR,LH
2008	MT,LR,LH
2009	MT,LR,LH

1-Itemsets	Sup-count
MT	5
MR	1
LR	5
LH	5

Prune

C1

Frequent Itemsets	Sup-count
MT	5
LR	5
LH	5

Join

L1

2-Itemsets	Sup-count
MT,LR	5
MT,LH	5
LR, LH	5

Prune

Frequent 2-Itemsets	Sup-Count
MT, LR	5
MT, LH	5
LR, LH	5

Join

Frequent 3-Itemsets	Sup-count
MT, LR, LH	5

Applying Association Rule:

Temperature (X,'13---29') ^ Rainfall (X,'0---10') ^ Humidity (X,'83---87' & X,'85---88') ⟹
Production (X,'3k---270k') where k=1000

April Month

MR=medium rainfall MR range=10---20
HR=high rainfall HR rang =20---30
LH=low humidity LH range= morning(83-87%) and afternoon(85-88%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	MT,HR,LH
2006	MT,MR,LH
2007	MT,HR,LH
2008	MT,MR,LH
2009	MT,MR,LH

1-Itemsets	Sup-count
MT	5
HR	2
MR	3
LH	5

Prune

Frequent Itemsets	Sup-count
MT	5
MR	3
LH	5

Join

2-Itemsets	Sup-count
MT,MR	3
MT,LH	5
MR, LH	3

Prune

Frequent 2-Itemsets	Sup-count
MT, MR	3
MT, LH	5
MR, LH	3

Join

Frequent 3-Itemsets	Sup-count
MT, MR, LH	3

Applying Association Rule:

Temperature (X,'13---29') ^ Rainfall (X,'10---20') ^ Humidity (X,'83---87' & X,'85---88') ⇨
Production (X,'271k---399k') where k=1000

May Month

HT=high temperature HT range=20°--36°C
MR=medium rainfall MR range=10---20
HR=high rainfall HR rang =20---30
MH=low humidity MH range= morning(88-92%) and afternoon(88-92%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	MT,HR,LH
2006	MT,MR,LH
2007	MT,HR,LH
2008	MT,MR,LH
2009	MT,MR,LH

1-Itemsets	Sup-count
HT	5
MR	1
MH	5
HR	4

Prune

Frequent Itemsets	Sup-count
HT	5
MH	5
HR	4

Join

2 Itemsets	Sup-count
HT, MH	5
HT, HR	4
MH, HR	4

Prune

Frequent 2-Itemsets	Sup-count
HT, MH	5
HT, HR	4
MH, HR	4

Join

Frequent 3-Itemsets	Sup-count
HT, HR, MH	4

Applying Association Rule:

Temperature (X,'20---36') ^ Rainfall (X,'20---30') ^ Humidity (X,'88---92' & X,'88---92') ⟹
Production (X,'400k---499k') where k=1000

June Month

HT=high temperature HT range=20°--36°C
MR=medium rainfall MR range=10---20
HR=high rainfall HR rang =20---30
MH=low humidity MH range= morning(88-92%) and afternoon(88-92%)
HH=high humidity HH range= morning(93-97%) and afternoon(90-94%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	HT,MR,HH
2006	HT,HR,HH
2007	HT,HR,MH
2008	HT,HR,HH
2009	HT,HR,HH

1-Itemsets	Sup-count
HT	5
MR	1
MH	1
HH	4
HR	4

Prune

Frequent Itemsets	Sup-count
HT	5
HH	4
HR	4

Join

2-Itemsets	Sup-count
HT, HH	4
HT, HR	4
HH, HR	4

Prune

Frequent 2-Itemsets	Sup-count
HT, HH	4
HT, HR	4
HH, HR	4

Join

Frequent 3-Itemsets	Sup-count
HT, HR, HH	4

Applying Association Rule:

Temperature (X,'20---36') ^ Rainfall (X,'20---30') ^ Humidity (X,'93---97' & X,'90---94') ⇨
Production (X,'500k---634k') where k=1000

July Month

HR=high rainfall HR rang =20---30

HH=high humidity HH range= morning(93-97%) and afternoon(90-94%)

Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	HT,HR,HH
2006	HT,HR,HH
2007	HT,HR,HH
2008	HT,HR,HH
2009	HT,HR,HH

1-Itemsets	Sup-count
HT	5
HH	5
HR	5

Prune

Frequent Itemsets	Sup-count
HT	5
HH	5
HR	5

Join

2-Itemsets	
HT, HH	5
HT, HR	5
HH, HR	5

Sup-count

Prune

Frequent 3-Itemsets	Sup-count
HT, HR, HH	5

Join

Frequent 2-Itemsets	Sup-count
HT, HH	5
HT, HR	5
HH, HR	5

Applying Association Rule:

Temperature (X,'20---36') ^ Rainfall (X,'20---30') ^ Humidity (X,'93---97' & X,'90---94') ⇨
Production (X,'635k---851k') where k=1000

August Month
HT=high temperature HT range=20°--36°C
LR=low rainfall LR range=0--10
HR=high rainfall HR range =20---30
HH=high humidity HH range= morning(93-97%) and afternoon(90-94%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item Ids
2005	HT,HR,HH
2006	HT,LR,HH
2007	HT,HR,HH
2008	HT,HR,HH
2009	HT,HR,HH

1-Itemsets	Sup-count
HT	5
HR	4
LR	1
HH	5

Prune

Frequent Itemsets	Sup-count
HT	5
HR	4
HH	5

Join

2 - Itemsets	Sup-count
HT, HR	4
HT, HH	5
HR,HH	4

Prune

Frequent 2-Itemsets	Sup-count
HT, HR	4
HT, HH	5
HR, HH	4

Join

Frequent 3-Itemsets	Sup-count
HT, HR, HH	4

Applying Association Rule:

Temperature (X,'20---36') ^ Rainfall (X,'20---30') ^ Humidity (X,'93---97' & X,'90---94') ⟹
Production (X,'800k---962k') where k=1000

September Month

HT=high temperature HT range=20°--36°C
LR=low rainfall LR range =0-10
HR=high rainfall HR range=20--30
MR=medium rainfall MR range=10--20
HH=high humidity HH range= morning(93-97%) and afternoon(90-94%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item Ids
2005	HT,LR,HH
2006	HT,LR,HH
2007	HT,HR,HH
2008	HT,LR,HH
2009	HT,MR,HH

1-Itemsets	Sup-count
HT	5
LR	3
HR	1
MR	1
HH	5

Prune

Frequent Itemsets	Sup-count
HT	5
LR	3
HH	5

Join

2-Itemsets	Sup-count
HT, LR	3
HT, HH	5
LR,HH	3

Prune

Frequent 3-Itemsets	Sup-count
HT, LR, HH	3

Join

Frequent 2-Itemsets	Sup-count
HT, LR	3
HT, HH	5
LR, HH	3

Temperature (X,'20---36') ^ Rainfall (X,'0---10') ^ Humidity (X,'93---97' & X,'90---94') ⇒
Production (X,'770k---938k') where k=1000

October month

HT=high temperature HT range=20°--36°C
LR=low rainfall LR range=0---10
MR=medium rainfall MR range=10--20
MH=medium humidity MH range= morning(88-92%) and afternoon(88-92%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	HT,MR,MH
2006	HT,LR,MH
2007	HT,LR,MH
2008	HT,LR,MH
2009	HT,MR,MH

1-Itemsets	Sup-count
HT	5
MR	2
LR	3
MH	5

Prune

Frequent Itemsets	Sup-count
HT	5
LR	3
MH	5

Join

2-Itemsets	
HT, LR	3
HT, MH	5
LR,MH	3

Sup-count

Prune

Frequent 3-Itemsets	Sup-count
HT, LR, MH	3

Join

Frequent 2-Itemsets	Sup-count
HT, LR	3
HT, MH	5
LR, MH	3

Applying Association Rule:

Temperature (X,'20---36') ^ Rainfall (X,'0---10') ^ Humidity (X,'88---92' & X,'88---92') ⇨
Production (X,'812k---925k') where k=1000

November Month
MT=medium temperature MT range=13°--29°C
LR=low rainfall LR range =0-10
MH=medium humidity MH range= morning(88-92%) and afternoon(88-92%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	MT,LR,MH
2006	MT,LR,MH
2007	MT,LR,MH
2008	MT,LR,MH
2009	MT,LR,MH

1-Itemsets	Sup-count
MT	5
LR	5
MH	5

Prune

Frequent Itemsets	Sup-count
MT	5
LR	5
MH	5

Join

2-Itemsets	Sup-count
MT, LR	5
MT,MH	5
LR,MH	5

Prune

Frequent 2-Itemsets	Sup-count
MT, LR	5
MT, MH	5
LR, MH	5

Join

Frequent 3-Itemsets	Sup-count
MT, LR, MH	5

Applying Association Rule:

Temperature (X,'13---29') ^ Rainfall (X,'0---10') ^ Humidity (X,'88---92' & X,'88---92') ⇒
Production (X,'424k---524k') where k=1000

December Month

LT=Low temperature LT range=9°--25°C
LR=low rainfall LR range =0-10
LH=low humidity LH range= morning(83-87%) and afternoon(85-88%)
Temperature width=16, Rainfall width=10, Humidity width=5

TID	List of item IDs
2005	LT,LR,LH
2006	LT,LR,LH
2007	LT,LR,LH
2008	LT,LR,LH
2009	LT,LR,LH

1-Itemsets	Sup-count
LT	5
LR	5
LH	5

Prune

Frequent Itemsets	Sup-count
LT	5
LR	5
LH	5

Join

2-Itemsets	Sup-count
LT, LR	5
LT, LH	5
LR,LH	5

Prune

Frequent 2-Itemsets	Sup-count
LT, LR	5
LT, LH	5
LR, LH	5

Join

Frequent 3-Itemsets	Sup-count
LT, LR, LH	5

Applying Association Rule:

Temperature (X,'9-25---36') ^ Rainfall (X,'0---10') ^ Humidity (X,'83---87' & X,'85---88') ⇨
Production (X,'80k---127k') where k=1000

4. CONCLUSION

The purpose of this paper was to obtain the information about production of Tea and explore which factors impacted on the production of Tea. Our results show that most of the tea productions have dependent rainfall, humidity, temperature, men power used, sunshine hours and type of seed used. From the factors we have seen that production is low in month of January, February and in the month of November and December. From the collected data it has been observed that at that time temperature is low (9°--25°C), rainfall is also less (0—10cm) and humidity is also low (morning (83-87%) and afternoon (85-88%)).Therefore production is comparatively low. After applying association rule we can find that productions nill.In the month of February temperature is increases from low to medium (13°--29°C).Tea is produced very low amounts. In the month of march temperature is medium (13°--29°C) and rainfall is medium(10-20cm) and it gradually increases production. In the month of December temperature is low (9°--25°C), rainfall is also less (0—10cm) and humidity is also low (morning (83-87%) and afternoon (85-88%)),therefore production is also low. In the .month of April temperature is medium(13°--29°C), rainfall is also medium(10-20cm)but humidity is low. Here the production is comparatively more then march. In the month of May temperature is high (20°--36°C),rainfall is also high(20-30cm) but the humidity is medium morning (88-92%) and afternoon (88-92%) . In this month the production is again increased.

In the month of June, July and August temperature is high(20°--36°C),rainfall is high(20-30cm)and humidity is also high morning (93-97%) and afternoon (90-94%)).Production in these months is maximum. Again in the data if we see in the month September temperature is high(20°--36°C),rainfall is low(0-10cm) and humidity is also high morning (93-97%) and afternoon (90-94%)).Therefore production is slightly decreased. October temperature is high(20°--36°C), rainfall is low(0-10cm)and humidity is medium morning (88-92%) and afternoon (88-92%) ,therefore production is further decreased. In November temperature is medium(13°--29°C), rainfall is low(0-10cm)and the humidity is also medium morning (88-92%) and afternoon (88-92%) From the statistical studies it is also cleared the same .From this study we can conclude that production of Tea is high in generally at high temperature, high rainfall and high humidity. But there are lots of other factors such as no of shade tree in cultivation, drain system, day length, tea pest control ,manuring,plucking and type of seeds used in the cultivation.

The correlation and association[7] amongst the factors are found after the analysis the knowledge be can used in decision making purpose for part of the organization which are directly or indirectly associated on the production of a Tea Garden.

The well use of the data mining tools in the tea cultivation and tea industry should bring revolutionary impact to the field. The study of tea cultivation processes is heavily based on the identification of understandable patterns which are present in the data. These patterns may be used for study factors affecting tea production. Data mining is at the care of the pattern recognition[1] process. Biologist and computing professionals should collaborate so that the two fields can contribute to each other. The challenge is for each to widen its focus to attain harmonious and productive collaboration to develop the best practices.

5. ACKNOWLEDGEMENT

The well use of the data mining tools[2] in the tea cultivation and tea industry should bring revolutionary impact to the field. The study of tea cultivation processes is heavily based on the identification of understandable patterns[4] which are present in the data. These patterns may be used for study factors affecting tea production. Data mining is at the care of the pattern recognition[1] process. Biologist and computing professionals should collaborate so that the two fields can contribute to each other. The challenge is for each to widen its focus to attain harmonious and productive collaboration to develop the best practices.

6. REFERENCES

[1] J. T. Tou and R. C. Gonzalez, "Pattern recognition principles," Addison-Wesley, London, 1974.

[2] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, "Data mining: A knowledge discovery approach," Springer, New York, 2007.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994

[4] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification,"Wiley,2001.

[5] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," Springer, New York, 2001.

[6] Aggarwal Charu and Yu Philip.Mining large itemsets for association rules.Bulletin Of the IEEE Computer Society Technical Committee on Data Engineering,21,no.1,March 1998

[7] Toivonen H.,Klemettinen M.,Ronkainen P.,Hatonen K and Mannila H.”Pruning and grouping discovered association rules”.Workshop on Statistical Machine Learning and Knowlege Discovery in Databases.1995

[8] Arun K Pujari ”Data Mining Techniques”Universities Press,Pages 69-109,February 2001

[9] Alex Beyson and Steve Smith , ”Data Warehousing,Data Mining and OLAP” 2004