Using Improved Agglomerative Hierarchical Technique to Rationalize Investment Decisions in Iraqi Banks

Ebtehal Talib Khudair

Ministry of Higher Education and Scientific Research

E_mail: talibebtehal@gmail.com

Bahgdad-Iraq

Abstract

The banking system relied on many decisions and is usually made manually, because of the large volume of data, the long time it takes to make a decision, and the possibility of a major error. Banks are now turning to data mining tools to predict payment delays, detect illegal transactions, record and approve credits, and more. In this paper, the researcher adopts the Improved Agglomerative Hierarchical Clustering of the classification of ethnic banks based on their annual profits for the period (2012-2015), to enable the investor to make the decision to invest his money in a bank that achieves great profits. The original algorithm was compared with the improved algorithm, and the results showed an improvement in execution time by 35 %.

Keywords: Banking Sector, Data Mining, and Agglomerative Hierarchical Algorithm

استخدام الأسلوب الهرمي التجميعي المحسن لترشيد قرارات الاستثمار في البنوك العراقية

ابتهال طالب خضير

وزارة التعليم العالي والبحث العلمي

بغداد–العراق

الخلاصـة

يعتمد النظام المصرفي على العديد من القرارات وعادة ما يتم اتخاذ تلك القرارات يدويًا، وبسبب الحجم الكبير للبيانات، والوقت الطويل الذي يستغرقه المدراء في اتخاذ القرارات، واحتمال حدوث خطأ كبير، بدأت تتحول المصارف إلى استخدام أدوات تنقيب البيانات لغرض التنبؤ بتأخير الدفع، واكتشاف المعاملات غير القانونية، وتسجيل الائتمانات والموافقة عليها، وأكثر من ذلك. في هذا البحث قام الباحث باستخدام خوارزمية التكتل الهرمية المحسنة لتصنيف المصارف العراقية بناءً على أرباحها السنوية للفترة (2012–2015)، لتمكين المستثمر من اتخاذ قرار استثمار أمواله في المصرف الذي يحقق ارباح كبيرة. تمت مقارنة الخوارزمية التكتلية لهرمية الأصلية مع الخوارزمية المحسنة، وأظهرت النتائج تحسنًا في وقت التنفيذ بنسبة 35%.

الكلمات المفتاحية: القطاع المصرفي وتنقيب البيانات والخوارزمية الهرمية التكتلية

Introduction

The banking industry is very competitive (chung, 2011), and the great development in technology has a clear impact on the banking industry (Rajab and Cumaar, 2017). The reason behind the concept of central databases is the wide spread of bank branches in different locations and thus the available data. Therefore, we must collect and manage the vast amount of data to be organized to build a picture of the environment of the bank and to solve any business problem (Rajab and Cumaar, 2017). In recent years the need to automatically extract knowledge has increased and the importance of speed of decision-making (Rajab and Cumaar, 2017). If decisions are made manually, the process takes time and effort and will not be a relatively accurate decision. Therefore, the solution will be towards electronic data management and analysis to extract the results and thus make the appropriate decision using data mining techniques (Preethi and Vijay, 2017). Technological development has enabled the banking sector to confrontation the challenges posed by the economy (Chitra and Subashini, 2013). At present, banks have realized their main tasks namely (customer retention, fraud detection, marketing, risk management, money laundering detection, as well as investment banking).

Data Mining Algorithms

Many problems in the banking sector can be solved by using data mining algorithms, generally are divided into two types: (Chitra and Subashini, 2013).

Classification Algorithms

This type of algorithm is defined as guided learning. Learning is guided by a pre-defined goal or trait. Its aim is to obtain predictive models. For example, if we want to promote a particular product to a number of customers, the model should be trained in the characteristics of customers who have already responded or did not respond to the promotional offer and thus can predict the category of customers we expect to respond to the offer (Chitra and Subashini, 2013).

Clustering Algorithms

This type of algorithm is defined as unsupervised learning. In this type there is no difference between independent and non-independent attributes to guide the algorithm in constructing a model. Non-directed learning is used for descriptive purposes and pattern detection. (Chitra and Subashini, 2013).

Hierarchical Clustering

This algorithm combines data elements to be a tree of groups. When we have a set of elements (N) to be grouped, we follow a set of steps for hierarchy. In the first step we place each element in a group, so the number of groups will be equal to the number of elements, and each group has one element. In the second step, the elements of the two most similar groups are merged into one group. The third step is to calculate the distances between the new group and the old groups. These steps (second and third) are then repeated until all elements are grouped. Depending on the distribution method, the hierarchy is divided into two types: Agglomerative and Divisive (Han, *et al*., 2012).

Agglomerative Hierarchical Clustering

The bottom-up approach is the basis of this algorithm. This algorithm places each of the elements in a single cluster, then begins to compare the elements among them, choose the two closest elements (based on a measure of similarity) and place them in a single cluster (ie, merge the most similar elements). Until all elements are placed in a single cluster or put a condition to

stop, for example, determine the number of final clusters show algorithm (1) (Prabha, *et al*., 2014).

Algorithm (1): Agglomerative Hierarchical Algorithm

Input X = (A1, A2, xA3, ..., An) #set of n vectors with several attributes
Output: B: one cluster
Step 1: Calculate the proximity matrix that contains the distance between each pair of patterns. Each pattern is a cluster.

Step 2: Find the most similar cluster pair # using cosine similarity distances (Eq.1).

$$similarity = \frac{A*B}{\|A\|_2\|B\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Step 3: Combine these two clusters into one cluster.
Step 4: Refresh the proximity matrix to reflect this merge.
Step 5: If there is one cluster, stop, else go to step 2.

Divisive Hierarchical Clustering

The descending approach is the basis of this algorithm. This algorithm places all elements in a single cluster and then divides the primary group into subgroups (based on similarity measures) and continues to divide gradually until place each element in a single group or determine the number of final groups as a stop condition (Prabha, *et al*., 2014).

Data Mining Applications in Banking Sector

Data Mining Applications in Banking Sector

At present, customers have opinions and trends in choosing the optimal bank to invest their money, so it is necessary that officials in the banking sector are able to attract the customer to invest and give them their attention because it is easy for customers to go to another bank (Chitra and Subashini, 2013). Therefore, the CRM strategy helps banks understand and determine the needs of their customers as well as create value for the brand by providing important information to customers in a timely manner. CRM systems, through dynamic customer information, provide the right services to the customer at the right time, as well as the possibility of building strong relationships with profitable customers. This system has three important phases in which mining is useful for the data: acquisition of the customer, increase its market value, and retention. The K-means aggregation algorithm is one of the most common algorithms used to arrange and group clients, for example, a bank divides customers according to their career into groups to provide the services they need. Other banks use Apriori algorithm to analyze the market basket by identifying groups of similar elements within a group (B) (Chitra and Subashini, 2013).

Fraud Detection in Banking Sector

Bank fraud is a concern for officials in many banks, and data mining techniques are among the most important tools used to detect bank fraud using the bank's customer databases (Bhasin, 2006). Fraud detection is the process of distinguishing real transactions from fraud and usually fraudulent through credit cards, and that the various frauds lose banks annually large sums of money so helps detect fraud early to reduce the losses of the bank. It is possible to study the behavior of the credit card user using the clustering technique and when looking at the results are extreme values that detect fraud, or through the probability of user behavior of the card between the past and present, and when you notice any significant deviation from the usual behavior of the customer the system generates alerts. The following are the most important examples of methods of bank fraud, including (credit

card, ATM, deposits, loans, Internet banking) (Rajab and Cumaar, 2017).

Marketing

Customer behavior can be predicted by identifying the current customer demand and comparing it with previous trends and the various services provided by the Bank using data mining techniques (Rajab and Cumaar, 2017). As well as the possibility of categorizing customers into categories (profitable and non-profitable) or by providing an attractive offer to customers by studying their interests to additional service or buy a product and this is called cross-selling (Chitra and Subashini, 2013). Cross selling is defined as a marketing concept through the use of data mining techniques. Therefore, the Bank provides financial services and products that enable it to retain its customers. (Chitra and Subashini, 2013). As for the retail sector, it relies heavily on data mining techniques in order to make the appropriate decision to sell products or provide services or grant incentives and promotions to customers. (Bhasin, 2006).

Risk Management

Banks risk their capital at each lending to a customer, and to assist the credit manager in making an appropriate decision or reduce the risk ratio, data mining techniques can be used to predict the behavior of the customer and his ability to repay the loan or slow the repayment or the inability of the client to repay the loan. By studying the historical data of the customer and knowing his ability to repay previous loans, he can predict his future behavior (Farooqi and Iqbal, 2017). The most important indicators to consider when analyzing customer data are sales trends, behavioral patterns, balance sheet figures, and verification of return patterns (Costa, *et al*., 2007).

Money Laundering Detection

The process of concealing the origin of illegal funds is money laundering in order to regulate it. Banks are a channel used to launder money, so governments and financial regulators often instruct banks to detect and prevent such transactions. Data mining techniques, especially the decision tree, are often used to reduce erroneous patterns and detect money laundering transactions (Pulakkazhy, *et al*., 2013).

Investment Banking

Investment in banks is usually for profit, so banks make offers to their clients through investment services to attract them. And based on the mining techniques, especially (K-means) technique for data clustering to provide the best services and offers to the customer through the analysis of customer files. As well as possible to provide the best investments to the client through an analysis of financial applications during time periods (Rajab and Cumaar, 2017).

Portfolio Management

In order to maximize profit, minimize losses and reduce risk, and by relying on data mining techniques, the portfolio can be managed well (chung, 2011). These techniques enable us to predict the prices of financial assets as well as the expected returns. Stocks for example, as well as the use of linear regression techniques and neural networks (Prabha, *et al*., 2014).

The Improved Agglomerative Hierarchical Algorithm

The similarity measure can be defined as the distance between different data points. While the similarity is the amount that reflects the strength of the relationship between two elements of the data (Singh, *et al*., 2013). In fact, the performance of many algorithms

depends on the choice of a good distance function across the input data set (Singh, *et al*., 2013). In this paper, for the purpose of improving the performance of the Agglomerative Hierarchical Algorithm, will replace the cosine distance measure with the Manhattan distance. Manhattan distance is a distance metric that calculates the absolute differences between coordinates of pair of data objects as shown in equation (2) given below (Singh, *et al*., 2013), algorithm (2) shows the implementation of Agglomerative Hierarchical Algorithm using Manhattan distance metric (Singh, *et al*., 2013)

$$Dist((x1,y1),(x2,y2)) = |x1 - x2| + |y1 - y2| \quad \ldots\ldots(2)$$

Algorithm (2): Improved Agglomerative Hierarchical Algorithm

Input X = (A1, A2, xA3, ..., An) #set of n vectors with several attributes.

Output: B: one cluster

Step 1: Calculate the proximity matrix that contains the distance between each pair of patterns. Each pattern is a cluster.

Step 2: Find the most similar cluster pair using the Manhattan_distance (Eq. 2)

$$Dist((x1,y1),(x2,y2)) = |x1 - x2| + |y1 - y2| \quad \ldots\ldots(2)$$

Step 3: Combine these two clusters into one cluster.

Step 4: Refresh the proximity matrix to reflect this merge.

Step 5: If there is one cluster, stop, else go to step 2.

The Implementation of Agglomerative Hierarchical Clustering and The Improved Algorithm

The Implementation of Original and Improved Agglomerative Hierarchical Clustering Algorithm on Iris Data Set

This improvement is represented by replacing the cosine distance measure with Manhattan distance. The results of proposed algorithm are compared with the classical algorithm in terms of execution time. Table (1) shows the result of applying both classical and improved algorithm (using Manhattan distance) on iris dataset by using 120 items as training set and 30 items as testing set.

Table (1): Results of the Comparison Between the Classical Agglomerative Hierarchical Clustering Algorithm and the Improved Algorithm

Dataset Name		Iris
K		2
Classical Agglomerative Hierarchical	Time in sec.	0.37
Improved Agglomerative Hierarchical	Time in Sec.	0.24

The Implementation of Improved Algorithm on Data of some Iraqi Banks

After reviewing the semi-annual report (2017) and the directory of the joint stock companies listed for the 2016 financial year for the Iraq Stock Exchange, Data was used for a group of Iraqi banks (19 banks). Where the researchers worked on the application of the agglomerative hierarchical algorithm on the annual net profit for banks for the years from 2012 to 2015. And the annual net profit was represented in the Iraqi dinar currency (in billion units), as shown in Table (2) *. The numbers in Table (1) were rounded up for ease of handling.

Table (2) Annual Net Profit of Banks for the Period (2012-2015)

Bank Code	The Bank's Annual Net Profit			
	2012	2013	2014	2015
1	13	9	9	7
2	25	32	28	6
3	24	21	4	5
4	1	27	29	17
5	15	14	7	2
6	23	12	13	12
7	16	16	12	7
8	1	1	2	4
9	5	4	5	4
10	21	3	0	5
11	31	47	36	10
12	18	43	14	0
13	17	16	10	11
14	12	25	17	20
15	34	30	22	18
16	7	5	1	48
17	11	8	6	1
18	34	38	22	27
19	2	9	9	0

*Preparing the researchers by dependence on the semi-annual report (2017) and the directory of the joint stock companies listed for the 2016 financial year for the Iraq Stock Exchange

Results and Discussion

When applying the algorithm to the data of the Iraqi banks, the time taken to classify the data into two clusters was (1.67) seconds. It was compiled on the basis of similarities in the financial returns for the period of time (2012-2015). Where the percentage of banks in the first cluster was (11%), while the proportion of banks (89%) was in the second cluster of the total banks. This indicates that banks in Iraq have relatively close profits. Consequently, the investor can make a better decision to invest his money in a bank that achieves the largest percentage of profits and moves away from banks with relatively few profits. Figure (1) shows the annual profits of Iraqi banks for the period (2012-2015).

Conclusion

The process of extracting knowledge from the available data is of great importance in the banking sector. In this paper, we reviewed a set of banking fields in which data mining operations have had a major impact on their development and growth, and among these fields (Customer Retention, Fraud Detection, Marketing, Risk Management, Money Laundering Detection, Investment, Portfolio Management). In this research, replacing the cosine distance scale in the classic hierarchical algorithm with Manhattan distance, and the results proved to reduce the decision-making time by a good rate. As well as discussed this research a practical application on the data of a group of Iraqi banks for the purpose of rationalizing the decision of the investor and directing his decision towards the best investment through the use of an Agglomerative Hierarchical Algorithm that which classified the banks according to the profits attributable to them in a very short time, and also showed that the banks are close to their financial returns.

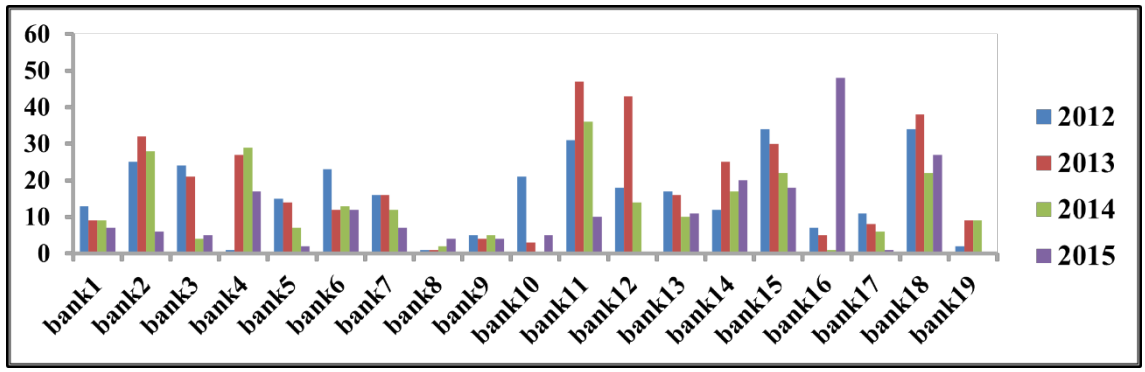

Figure (1): The Annual Profits of Iraqi Banks for the Period (2012-2015)

References

Bhasin, M. L. (2006). Data mining: A competitive tool in the banking and Retail Industries, The Chartered Accountant 588- 594.

Chitra, K. & Subashini, B. (2013) An Efficient Algorithm for Detecting Credit Card Frauds, Proceedings of State Level Seminar on Emerging Trends in Banking Industry, March 2013.

Chitra, K. & Subashini, B. (2013) Data Mining Techniques and its Applications in Banking Sector, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 8.

Costa, G. F.; Folino, A.; Locane, G.; Manco and R. Ortale (2007) Data Mining for Effective Risk Analysis in a Bank Intelligence Scenario. Preccedings of the 23rd International Conference on Data Engineering Workshop, Apr. 17-20, IEEE Xplore Press, Istanbul, pp, 904-911.

Farooqi, Md. Rashid & Iqbal, Naiyar (2017) Effectiveness of Data Mining in Banking Industry: An Empirical Study, International Journal of Advanced Research in Computer Science, 8(5).

Han, J.; Pei, J. and Kamber, M. (2012) "Data Mining Concepts and Techniques" Third Edition, Morgan Kaufmann Publications, Printed in the United States of America.

Prabha, S.; Duraiswamy, K. and Sharmila, M. (2014) "Analysis of Different Clustering Techniques in Data and Text Mining" International Journal of Computer Science Engineering (IJCSE), 3(2).

Preethi, M. and Vijay alakshmi, M. (2017) Data Mining in Banking Sector, International Journal of Advanced Networking & Applications (IJANA), 8(5), pp, 1-4.

Pulakkazhy, S. and Balan, R. V. S. (2013) Data Mining in Banking and Its Applications- A Review, Journal of Computer Science 9 (10), pp, 1252-1259.

Rajab H. C. and Cumaar, S. Y. (2017) a Study of Data Mining Applications in Banking, International Journal of Pure and Applied Mathematics, 116(15), pp, 265-271.

Singh, A.; Yadav, A. and Rana, Ajay (2013) K-means with Three Different Distance Metrics, International Journal of Computer Applications (0975 – 8887), 67(10).

Tak-chung, F. (2011) A review on Time Series Data Mining, Eng. Appli. Artif. Intell., 24, 164-181.