IMPORTANT BUSSINESS FACTOR ANALYSIS USING DATAMINING APPROACH IN FINANCIAL SECTOR

Article · February 2012

CITATIONS
0

READS
47

1 author:

Annamalai Suresh
Nehru Group of Institutions
45 PUBLICATIONS **132** CITATIONS

SEE PROFILE

IMPORTANT BUSSINESS FACTOR ANALYSIS USING DATAMINING APPROACH IN FINANCIAL SECTOR

A.SURESH

Research scholar,
Dr.M.G.R University-Chennai.
Email: prisu6esh@yahoo.com

Dr.K.L.SHANMUGHANATHAN

Prof & Head, Dept of CSE,
R.M.K Engg College-Chennai.
Email: kls_nathan@yahoo.com

Abstract

The term cluster analysis encompasses a number of various types of algorithms and methods for grouping objects of similar kind into relevant categories. A common question facing researchers in a lot of areas of analysis is how to organize experimental data into meaningful structures, that is, to develop taxonomies. In other words cluster analysis is an examining data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the similar group and minimal otherwise. Given the above, cluster analysis can be used to determine structures in data without providing clarification/interpretation. In other way, cluster analysis simply discovers structures in data without explaining why they exist.

Introduction

Data Mining is an analytic process deliberate to investigate data (usually large amounts of data - typically business or market related) in search of dependable patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data [1]. The definitive goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications [2, 3]. The progression of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern classification with validation/verification, and (3) deployment (i.e., the application of the representation to new data in order to generate predictions).

Connotation of Testing

In fact, cluster analysis is not as much a distinctive statistical test as it is a "collected works" of different algorithms that "put objects into clusters according to well defined resemblance rules." The point here is that, unlike many other statistical measures, cluster analysis methods are mostly used when we do not have any a priori hypotheses, but are still in the exploratory phase of our research [4]. In a sense, cluster analysis finds the

"most significant solution possible." Therefore, statistical consequence testing is really not appropriate here, even in cases when p-levels are reported (as in *k*-means clustering).

Various Type of Clustering

Well Separated

A set of objects in which each object is closer to every object in the cluster than to any object not in the cluster. Sometimes a threshold is used to stipulate that all the objects in a cluster must be adequately close to another.

Prototype

A set of objects in which each object is closer to the prototype that defines the cluster than to the prototype of any other cluster either data with continuous attributes, the prototype of a cluster is often a centroid, i.e., mean of all the points in the cluster. When a centroid is not evocative, such as when the data has definite attributes, the prototype is often a medoid.

Graph-Based

If the data is represented as a graph, where the nodes are objects and the links represent connections among objects than a cluster can be defined as a connected component the example of graph-based clusters are contiguity-based clusters, where two objects are connected only if they are within a specified distance of each other this implies that each object in a contiguity-based cluster is closer to some other object in the cluster to any point in a different cluster.

Shared-Property

Commonly, we can define a cluster as a set of objects that share some property. This definition encompasses all the previous definitions of a cluster. For example, objects in a center-based cluster share the property that they are all closest to the same centroid either medoid. But, the shared-property approach also includes new types of clusters.

SPACE AND TIME COMPLEXITY

The space necessities for k-means are modest since only the data points and centroid are stored. Particularly, the storage essential is o ((m+K) n), where m is the number of points and n is the number of attributes [5, 6]. The time necessities for K-means are also modest-basically linear in the number of data points. In particular, the time also required for o (I*K*m*n), where I is the number of iterations essential for convergence. As mentioned, I is often small and can generally be safely bounded, as most changes typically occur in the first few iteration. Hence, K-means is linear in m, the number of points, and is efficient as well as simple provided that K, the number of clusters, is notably less than m.

Minimal Clustering Approach

Step 1: Preclustering: Making Small Clusters

The first step of the two-step process is construction of preclusters. The goal of preclustering is to decrease the size of the matrix that contains distances between all probable pairs of cases. Preclusters are just clusters of the unique cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decide, based on a distance compute, if the current case should be merged with a previously shaped precluster or starts a new preclustering [7, 8]. When preclustering is comprehensive, all cases in the similar preclustering are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the no. of preclusters.

Step 2: Hierarchical Clustering of Preclusters

In the second step, the standard hierarchical clustering algorithm on the preclusters. A form cluster hierarchically let explore a range of solutions with dissimilar numbers of clusters.

Data Selection

Data mining in finance has the similar challenge as common data mining in data selection for building models. In this progress question is tightly connected to the selection of the target variable. There are different options for target variable y*: $y=T(k+1)$, $y=T(k+2)$…, $y=T(k+n)$,* where $y=T(k+1)$ represents estimate for the next time moment, and $y=T(k+n)$ represents estimate for *n* moments ahead. Selection of dataset ***T*** and its size for an exact favored n is a considerable challenge.

Relational Methodologies Based on Attributes

Different parameters distinguish data mining methodologies for the process. Data categories and mathematical algorithms are most essential among them. The first data type is representing by attributes of objects, which are each, object x is given by a set of values A1(x), A2(x)…… An (x). The common data mining methodology expect this type of data and it is known as an attribute-based or attribute-value methodology. It covers a large range of statistical and connectionist (neural network) methods. The relational data type is a second type, where objects are represent by their relations with other objects, for instance, x>y, y<z, x>z. In this example we may not know that x=3, y=1 and z=2. Thus attributes of objects are not known, but their relations are known already. Objects may have different attributes (e.g., x=5, y=2, and z= 4), but still have the same relations. Less conventional relational methodology is based on the relational data type. Another data distinctive important for financial modeling methodology is an actual set of attributes involved. A fundamental analysis approach incorporates all available attributes, however, technical analysis approach is based only on a time series such as stock price and parameters derived from it. Most popular time series are index value at open, index value at close, highest index value, lowest index value and trading volume and lagged returns from the time series of interest [9]. Primary factors include the price of gold, retail sales index, industrial production indices, and foreign currency exchange rates. Technical factors include variables that are resulting from time series such as moving averages [10]. The next characteristic of a specific data mining method is a form of the connection between objects. Many data mining methods assume efficient form of the relationship.

Risk Factor Analysis Using Data Mining Concept

Decision-making under assurance: In this case the decision maker has the complete knowledge of Consequence of every decision selection with certainty. Apparently will choose another that yields the largest return (payoff) for the known future. In this decision model, certainty means that only one possible state of nature (future) exists [11]. Decision-making under risk: In this case the decision-maker has less than the complete knowledge with assurance of the consequences of every decision choice. This means there is more than one state of nature and for which he /she makes an assumption of the probability with which each state of nature will occur [12,13]. Decision making under improbability: In this case the decision maker is unable to specify the probabilities with which the various states of nature will occur. Thus, decisions under uncertainty are taken with even less information than decisions under risk. Most of the business decisions are based on the future planning of business operations, depending on the timeframe and seasonality factors [14]. Data mining tools help in formulation of both deliberate and tactical decisions that are essential for an organization to survive in the business, therefore data mining and statistics tools can be regarded as the best solutions to decision making under risk and uncertainty [15, 16].

Time Series Analysis

The purpose of using time series analysis is to discover the components of time series namely Long term trend (T), Seasonal variations (S), cyclic variation (C) and Irregular movements (I). Thus a time series is represented by (T+C+S+I) call it as an additive model and (T×S×C×I) as multiplicative model and ((T×S) +(C×I)) as a mixed model. The basic econometric applications of time series analysis is the use of multiple regression technique in this, several independent variables are used to check the correctness of the prediction of dependent variable by its R –Square value. Suppose the dependent variable is the price of a particular stock then the independent variables can be market index, inflation, GDP, Index of Industrial Production (IIP) etc [17, 18]. In this several regression analysis one cannot conclude the results just by seeing good R-Square value, because higher the R-Square value may lead to the multicollinearity problem, which correspond to basically the existence of, inter correlations among independent variables that causes false results in the regression calculations. To face the problems of econometrics namely Multicollinearity, Heteroskedasticity and Auto correlation statisticians have developed so called time series econometrics. The basic application of these methods in finance includes EARMA (Extended Auto Regressive Moving Average) and ARIMA modeling techniques.

Conclusion

Data mining are apprehensive with learning from data and transformation of that data into functional information, data mining helps to find out the patterns and associations between the variables in the data and statistics helps to get processed that data into useful information. Since data mining approach covers empirical models and regularities derived directly from data with little domain knowledge explicitly involved. Previously, in most of the domains, deep field-specific theories emerge after the field accumulates enough empirical regularity. The future of data mining would be to create more empirical regularities and unite them with domain

knowledge via generic analytical data mining approach. Exploring the valuable data with data mining techniques and processing of such data by statistical software is gaining importance in today's business area. All national and MNC's are using various data mining techniques and analytics to explore the relationship between various micro and macro economic factors and make policies using data analytics. The importance of a new innovative approach to revive this sector becomes significant.

Reference

[1] Stephen PRobbins, Management science [M], Beijing: China Renmin University Press, 1997.
[2] Jiangtao Ren, Xiaoxiao Shi, etal, An Improved K-Means Clustering Algorithm Based on Feature Weighting[J], Computer Science, 2006, 33(7): 186-187.
[3] T M Mitchell, Machine Learning. New York: McGraw-Hill Companies Inc, 1997, 230-247.
[4] J Basak, R K De, S K Pal. Unsupervised feature selection using a neuro-fuzzy approach. Pattern Recognition Letters, 1998, 19(11): 997-1006.
[5] K. Wagstaff, C. Cardi, S. Rogers, and S. Schroedl, "Constrained K-means with Background Knowledge", Proceeding of the 18th International Conference on Machine Learning, pp. 577-584, 2001.
[6] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding", Proceeding of the 19th International Conference on Machine Learning, Sydney, Australia, pp. 19-26, 2002.
[7] M. Halkidi, D. Gunopulos, N. Kumar, M. Vazirgiannis, and C. Domeniconi, "A Framework for Semi-Supervised Learning based on Subjective and Objective Clustering Criteria", Proceedings of the IEEE Conference on Data Mining (ICDM), Nov. 2005.
[8] M. Ceccarelli, and A. Maratea, "Semi-supervised Fuzzy c-Means Clustering of Biological Data", Proceeding of WILF 2005, pp. 259-266, 2005.
[9] S. Koyuncugil, Fuzzy Data Mining and its application to capital markets. Ph.D. dissertation, Dept. Statistics, Ankara University, Ankara, Turkey, 2006.
[10] S. Koyuncugil. and N. Ozgulbas. Early Warning System for SMEs as a financial risk detector in Data Mining Applications for Empowering Knowledge Societies. Hakikur Rahman, Ed, Idea Group Inc., USA, 2008.
[11] Berson, S. Smith, and K. Thearling, Building Data Mining Applications for CRM. McGraw-Hill, USA. 2000.
[12] Ivakhnenko, A. G. & Madala, H. R., Inductive learning algorithms for complex systems modeling, CRC press, Inc, 1994.
[13] Kolari, J., Glennon, D., Shin, H., & Caputo, M.,"Predicting large US commercial bank failures", Journal of Economics and Business, Vol 54, No 321, pp.361-387, 2002.
[14] Lam, K. F., & Moy, J. W., "Combining discriminant methods in solving classification problems in two-group discriminant analysis", European journal of Operational Research, Vol 138, pp. 294-301, 2002.
[15] Lee, K., Booth, D., & Alam, P., "A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms", Expert Systems with Applications, Vol 29, pp. 1–6, 2005.
[16] Lee, K. C., Han, I., & Kwon, Y., "Hybrid neural network models for bankruptcy predictions", Decision Support Systems, Vol 18, pp. 63–72, 1996.
[17] T. Minka, "Bayesian inference, entropy, and the multinomial distribution," 2003 [Online].
[18] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 105–161.

Mr. A.Suresh B.E.,M.Tech.,(Ph.D) works as the Associate Professor & Head of the Computer Science and Engineering Department in ASAN Memorial College of Engineering & Technology, Chengalpet, Chennai, TamilNadu, India. He has more than 13 years of experience in teaching and his areas of specializations are Data Mining, Artificial Intelligence, and System Software. He has published many of his research work in national and international conference and he has published one book in the name of Data structures & Algorithms in DD Publications.

Dr. K.L. Shunmuganathan B.E,M.E.,M.S.,Ph.D works as the Professor & Head of CSE Department of RMK Engineering College, Chennai, TamilNadu, India. He has more than18 years of teaching experience and his areas of specializations are Networks, Artificial Intelligence, and DBMS.