Data Mining in Marketing: Framework and Application to Supply Chain Management

Steven H. Kim and Sung Hwan Min

Graduate School of Management
Korea Advanced Institute of Science and Technology
Seoul, Korea
Email: skim@msd.kaist.ac.kr
Web: http://kgsm1.kaist.ac.kr/dmm

Abstract

The objective of knowledge discovery and data mining lies in the generation of useful insights from a store of data. This paper presents a framework for knowledge mining to provide a systematic approach to the selection and deployment of tools for automated learning. Every methodology has its strengths and limitations. Consequently, a multistrategy approach may be required to take advantage of the strengths of disparate technique while circumventing their individual limitations. For concreteness, the general framework for data mining in marketing is examined in the context of developing agents for optimizing a supply chain network.

Key words: Knowledge discovery, data mining, multistrategy learning, marketing, supply chain management, agents.

PURPOSE

The objective of knowledge discovery and data mining is to support decision making through the effective use of information. The practical implications for knowledge mining lie in the development of learning software to discover patterns, trends or relations in databases (Fayyad, 1996; Wimberley, 1995).

The automated approach to knowledge discovery is especially useful when dealing with large data sets or complex relationships. For many applications, automated software may find subtle patterns which escape the notice of manual analysis, or whose complexity exceeds the cognitive capabilities of humans. Secondly, software for data mining may ensure the practicality of analyzing large data sets which would require an unjustifiable amount of manual effort. Lastly, automated learning tools may monitor complex environments such as distribution networks on a continuous basis, updating their knowledge base immediately with the outbreak of new developments.

BACKGROUND

Context. The past few decades have witnessed increasing interest in the development of software for knowledge mining. The tools have been applied widely to practical domains especially since the late 1980s.

To date, however, practical applications have tended to utilize single techniques in isolation. However, each tool has its advantages and drawbacks. For this reason, a multistrategy approach to knowledge mining offers the potential to take advantage of the strengths of various techniques while bypassing the limitations of each methodology.

Methods. Neural nets offer many advantages such as robustness and graceful degradation (Hopfield, 1982; etc.). However, they suffer from a number of limitations such as the need for long learning times. Another drawback lies in the implicit nature of the acquired knowledge, which cannot be explicitly communicated to a human decision maker (Kim, 1994a, 1994b; etc.).

A key advantage of case based reasoning lies in the ability to work with data in their original format. Often the methodology is effective even when applied to an incomplete or partially faulty database. One drawback, however, lies in the tendency of conventional CBR tools to identify similarities based on superficial rather than substantive features of two cases. Another limitation is found in the ability to perform well- such as in predicting stock markets- without yielding an explicit explanation of the underlying causative factors (Kim and Kang, 1996; Kim and Novick, 1993). However, this is a limitation applicable to the entire gamut of learning tools.

Induction refers to the generation of rules or the classification of objects into decision trees. The methodology generates new knowledge in the explicit form of rules or trees (Quinlan, 1986; etc.). However, the methodology can be cumbersome to use and has not been deployed as widely as it deserves.

Statistical techniques represent a long-standing approach to acquiring insight into the nature of data (Kim, 1992; etc.). On the other hand, statistical methods such as regression, factor analysis, and other methods tend to focus on linear relationships. Consequently, their applicability to the nonlinear relations of reality remain bounded.

As indicated earlier, a multistrategy approach incorporates two or more elementary techniques. The resulting synergism can yield superior performance, as illustrated by an integrated system embodying both case reasoning and a neural network (Kim and Oh, 1996). A limitation of the multistrategy approach lies in the fact that the road to synergism is not always apparent at the current state of art. Even so, this dearth of experience will be rectified in the years to come.

Agents. Mobile agents are active, autonomous objects which roam through locations in a so-called agent system. A distributed abstraction layer provides security for the component systems as well as the concepts and mechanisms for mobility and communication. An example of a Java based mobile agent system is found in Mole (Strasser et al., 1996).

FRAMEWORK

A 2-dimensional framework for knowledge mining is presented in Table 1. The exhibit portrays two dimensions relating to function and technique. The functional dimension consists of the categories of prediction, classification, rule generation, decision trees, dependency relations, statistical summarization, and visualization. On the other hand, the dimension of technique comprises neural networks, case based reasoning (CBR), induction, statistics and multistrategy approaches.

Each entry in Table 1 lies in the intersection of two dimensions. For instance, orthogonal arrays represent a statistical technique which can be used for prediction.

A particular tool may embody more than one technique and/or function. To illustrate, MLC++ is a software package for induction which supports the tasks of both classification and visualization.

APPLICATION OF FRAMEWORK

The general framework for knowledge mining may be tailored to specific domains. For the sake of concreteness, this paper focuses on a particular field from the realm of marketing.

Table 2 presents examples of the relevance of knowledge mining tools to the realm of marketing. For instance, case based reasoning may be applied to the classification of designs, the configuration of distribution networks and the servicing of customer inquiries.

CASE STUDY

The supply chain of a large enterprise involves a world-wide network of suppliers, facilities, warehouses, distribution centers and retailers. In this network, raw materials are procured, transformed into products, delivered to customers, and serviced.

Supply chains have very complex multitiered structures, and are therefore too difficult to model through analytic formulas. On the other hand, an agent architecture represents an effective way to depict, analyze, and optimize supply chains.

A supply chain involves the sequence of functions relating to the acceptance of a customer's order, producing the desired product, and delivering it. In general, an enterprise is supported by these critical business processes: (1) product development, (2) order fulfillment, and (3) customer service. To remain responsive, the business processes should be monitored carefully and optimized continually in order to remain competitive in a rapidly shifting environment.

An order fulfillment process (OFP) begins with receipt of an order from a customer and ends with the delivery of the finished good. The cycle time for order fulfillment covers the period from the receipt of an order to the delivery. The process involves the coordination of diverse activities ranging from sales commitment, credit acceptance, manufacturing, logistics, accounts receivable, and interactions with external suppliers. The main activities of the OFP can be summarized as follows:

1. *Order management,* relating to the receipt and acceptance of an order.
2. *Manufacturing,* including production scheduling,

material requirements planning, capacity planning, and shop floor control.

3. *Distribution,* which involves logistics such as inventory, transportation, and so on.

A nested abstraction of a supply chain network (SCN) is depicted in Figure 1. The next figure portrays the order fulfillment process within a supply chain network.

Agility refers to the rapid accommodation of unpredictable changes. In business process redesign, agility involves the following dimensions:

1. Speed, or reduction of cycle times.
2. Flexibility, involving a large repertoire of potential responses.
3. Robustness, denoting the fulfillment of goals despite exceptional circumstances.
4. Adaptability, pertaining to the self-learning ability of the system.

An example of a multi-agent simulation package for modeling complex adaptive systems lies with Swarm (Minar et al., 1996). The software provides a general purpose tool for constructing simulation models.

Aglets are Java-based agents developed at IBM (Bigus. and Bigus, 1997). Aglets are not necessarily intelligent. However, since they are implemented in Java, they can be extended through adaptive components.

Table 3 presents a mapping of the properties between an agent system and Swarm. The Swarm platform can be employed to model supply chain networks. The mapping of properties between an SCN and Swarm is presented in the next table.

An OFP consists of two types of movements: material flow and information flow, as indicated in Figure 2. Consequently two kinds of agents are needed to model an OFP.

An *object agent* represents a physical object such as a firm, factory, production line, machine, and so on. In contrast, a *process agent* represents a set of functions or tasks. Process agents include the following.

- Demand forecasting agent: forecasting demand using data mining techniques.
- Order management agent: processing orders from external entities or other agents within the organization.
- Production planning agent: creating plans within different planning horizons.
- Inventory management agent: handling tasks involving inventory management.
- Capacity planning agent: monitoring and deploying capacity within a manufacturing system.
- Materials planning agent: securing materials needed for production.
- Shopfloor control agent: dispatching parts and providing capacity utilization information for other agents.
- Manufacturing agent: implementing production processes.
- Supply chain network management agent: coordinating suppliers and customers.

Figure 3 presents a system developed to model the OFP in an SCN based on Swarm. The OFP in the model governs the entire simulation model. The performance of the order fulfillment process is characterized by metrics such as the order cycle time, order fulfillment rate, inventory cost, tardiness on committed due dates, and customer expected lead time.

By incorporating data mining techniques within individual agents, it is possible to develop an agent system which learns autonomously to enhance system performance. For instance, a multistrategy technique may be embedded in each agent to accelerate the learning process. An example of such a strategy is presented in Figure 4. It presents a multistrategy procedure for optimizing case based reasoning through directed random search. These and other multistrategy techniques are currently under investigation in our lab.

CLOSURE

The objective of knowledge discovery and data mining lies in the generation of useful insights from a store of data. This paper presents a framework for knowledge mining to provide a systematic approach to the selection and deployment of tools for automated learning in the realm of marketing. In addition, a framework for data mining agents was illustrated in the context of optimizing a supply chain network. Supply chains have complex multitiered structures, and therefore are usually too difficult to model analytically. On the other hand, an agent approach is a promising approach to the analysis and design of an SCN.

A direction for future research is to validate the proposed framework through adaptive agents. The agents will incorporate multistrategy learning techniques such as the methodology illustrated in this paper.

REFERENCES

Anand, T., et al. "KDD-95 panel on Commercial KDD Applications: The 'Secret' Ingredients for Success." http://www-aig.jpl.nasa.gov/kdd95/KDD95-panels.html, 1995.

Baudin, M. "Introduction to AcknoSoft."

http://www.acknosoft.com/#Industrial-maintenance, 1996.

Bigus, J.P., Bigus, J. Constructing Intelligent Agents with Java, Willey Computer Publishing, 1997

Erkes, J. W., et al. "Implementing Shared Manufacturing Services on the World-Wide Web." *Communications of ACM,* v. 39(2), 1996: 34-45.

Fayyad, U. M., et al. *Advances in Knowledge Discovery and Data Mining*, AAAI Press / MIT Press, 1996.

Fayyad, U. M., et al. "Quest Data Mining Technologies." http://www.almaden.ibm.com/cs/quest/technologies.html, 1996.

Feldman, R., et al. "KDD95 Description of Demos." http://www-aig.jpl.nasa.gov/kdd95/KDD95-demo.html, 1996.

GTE. "Siftware: Tools for Data Mining and Knowledge Discovery." http//info.gte.com/~kdd/siftware.html, 1994.

Hardwick, M., et al. "Sharing Manufacturing Information in Virtual Enterprises." *Communications of ACM,* v. 39(2), 1996: 46-54.

Hopfield, J. J. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proc. Nat. Acad. Sciences USA,* v. 79(8), 1982: 2554-2558.

Kim, S. H. *Statistics and Decisions.* New York: Van Nostrand Reinhold, 1992.

Kim, S. H. "Learning Systems for Process Automation through Knowledge Integration." *Proc. Second World Congress on Expert Systems,* Lisbon, Portugal, 1994a: CD, item 63.

Kim, S. H. *Learning and Coordination.* Dordrecht, Netherlands: Kluwer, 1994b.

Kim, S. H. *Data Mining in Finance.* Seoul: Sigma Consulting Press, 1999.

Kim, S. H. and M.B. Novick. "Using Clustering Techniques to Support Case Reasoning." *International J. of Computer Applications in Technology,* v. 6(2/3), 1993: 57-73.

Lee, B. H. and Ishii, K. "Product-Agents and the Embedded Web: Architecture and Tools for Lifecycle Design." http://dfe.stanford.edu/embedagents/aid96.html, 1996.

Lin, F. , Tan, G.W. and Shaw, M.J. " Modeling Supply-Chain Networks by a Multi-Agent System." *Proc. 31st Annual Hawaii International Conference on System Sciences*, 1998.

Minar, N. , Burkhart, R., Langton, C., Askenazi, M. "The Swarm Simulation System: A Toolkit for Building Multi-Agent Simulations." http://www.santafe.edu/projects/swarm, 1996.

Nutek, Inc. "Qualitek-4 Software: Taguchi method, Design of Experiments, Reliability, Seminars and Software." http://www.wwnet.com/~rkroy/wp-q4w~1.html, 1996.

Peterson, G. E., et al. "Using Taguchi's Method of Experimental Design to Control Errors in Layered Perceptrons." *IEEE Transactions on Neural Networks,* v. 6(4), July, 1995: 949-961.

Petrie, C. J. "Agent-based Engineering, the Web, and Intelligence." http://cdr.Stanford.EDU/NextLink/Expert.html, 1996.

Phadke, M. S. *Quality Engineering Using Robust Design,* Prentice Hall, 1989.

Quinlan, J. R. "Induction of Decision Trees." *Machine Learning,* v. 1(1), 1986: 81-106.

Sasisekharan, R., V. Seshadri, and S. M. Weiss. "Data Mining and Forecasting in Large-Scale Telecommunication Networks." *IEEE Expert,* February, 1996: 37-43.

Silicon Graphics. "Machine Learning Library in C++." http://www.sgi.com/Technology/mlc/docs.html, 1996.

Strasser, M., Baumann, J., Hohl, F. "Mole - A Java Based Mobile Agent System." http://www.informatik.uni-stuttgart.de/ipvr/vs/Publications/Publications.html# 1996-strasser-01, 1996

Swami, A. "Data Mining with Silicon Graphics Technology." Silicon Graphics, Inc. http://www.sgi.com/software/mineset/mineset_data.html, 1995.

Vieira, J., et al. "CASTING: CAse-based reasoning: STimulation of INdustrial usaGe." MARI Computer Systems Limited, UK, 1994.

Watson, I. "Case-Based Reasoning Development Tools: A Review." University of Salford, Bridgewater Building salford, UK, 1996.

Wimberley. L. "Lockheed Corporation's Recon System." Lockheed Artificial Intelligence Center. http://hitchhiker.space.lockheed.com/~recon/data_mining.shtml, 1995.

Zhou, Q., et al. "A New Tool Life Criterion For Tool Condition Monitoring Using a Neural network." *Engineering Applications of Artificial Intelligence,* v. 8(5), 1995: 579-588.

Table 1. Typology of data mining tools by function and technique. The entries in the table illustrate several pertinent tools (Anand, 1995; Baudin, 1996; GTE, 1994; Kim, 1999; Nutec, 1996; Silicon Graphics, 1996; Swami, 1995; Vieira et al., 1994; Watson, 1996; Wimberley, 1995).

TECHNIQUE FUNCTION	Neural network	CBR	Induction	Genetic Algorithm	Statistics	Multistrategy
Prediction	• Neural Connection • Recon	• KATE	• KATE	• Generator • Ultragem	• Qualitec-4 (Orthogonal Arrays)	• Second order learning
Classification	• MATLAB Neural Network Toolbox • Inspect • Neural Connection	• ECOBWEB • KATE	• SE-learn • OC1 • MLC++ • DBlearn	• Ultragem • Evolver	• SPSS CHAID • Auto Class C (ACC)	• Inspect • Darwin • Clementine • Modelware Professional
Rule generation	• Recon	• ESTEEM • IDIS	• CN2 • Brute • QM	• Generator		• Recon
Decision trees		• KATE	• KATE • OC1		• Information Harvester	• Clementine
Dependency relations					• TETRAD II	
Statistical summaries	• 4Thought • Data Engine • Neural Connection	• Case Power		• Generator	• MATLAB • SAS • SPSS	• Darwin
Visualization		• Case Power	• MLC++	• Generator • Evolver	• Inspect • Clementine • Recon	

Table 2. Examples of data mining applications in marketing.

- Neural network
  - Market segmentation
  - Forecasting sales levels based on shipments
  - Identification of product families
- Case based reasoning
  - Grouping of product designs
  - Configuration of distribution networks
  - Help support for product information from potential customers
- Induction
  - Customer segmentation
  - Feature discretization
  - Trend forecasting
- Genetic algorithms
  - Resource optimization
  - Distribution planning
  - Logistics planning
- Statistics
  - Clustering of product groups
  - Forecasting production requirements
  - Parameter selection through orthogonal arrays
- Multistrategy
  - Synthesis of advertising strategies
  - Improvement of scheduling
  - Visual exploration of data sets
  - Prediction of nonlinear interactions

Table 3. The mapping of properties between mining agents and the Swarm system.

Data Mining Agent Systems	Swarm Simulation System
Agent	Object-oriented representation
Self-organization	Nested hierarchy
Autonomous agent behavior	Individual-based modeling
Distributed and loosely coupled control	Parallel and concurrent processing

Table 4. Mapping of properties between SCN and Swarm.

Supply Chain Network	Swarm Simulation System
Composition of autonomous and semi-autonomous business entities.	A swarm of agents with individual modeling.
Business entities perform different organizational roles.	Agents with internal state variables and action functions.
Multiple layer abstraction.	Nested hierarchy.
Information flow between business entities.	Message passing between agents.
Material flow during procurement, manufacturing, and distribution activities.	Discrete event simulation and time-stepped scheduling to trigger agent actions.
In an SCN, the activities of individual entities contribute to global performance.	In Swarm, the combination of individual behaviors determines collective behavior.
Visibility of an SCN is determined by the information boundary.	Boundaries of links for message passing determine visibility.

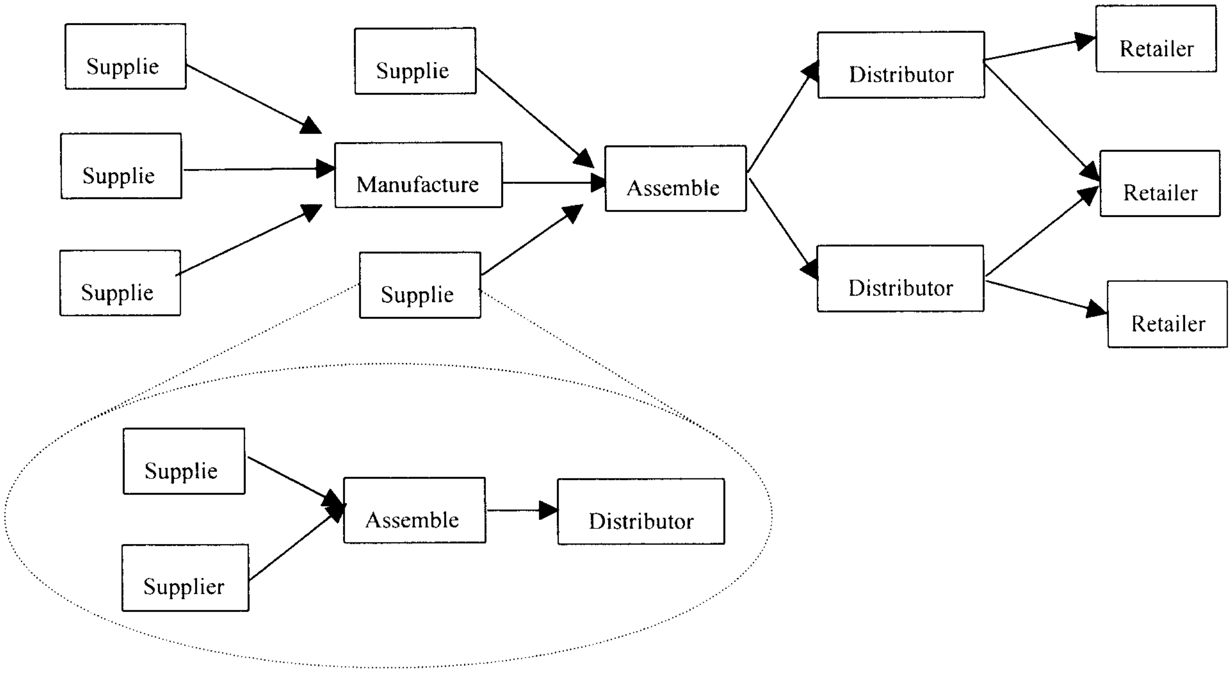


Figure 1. The nested abstraction of an SCN (Lin et al., 1998)

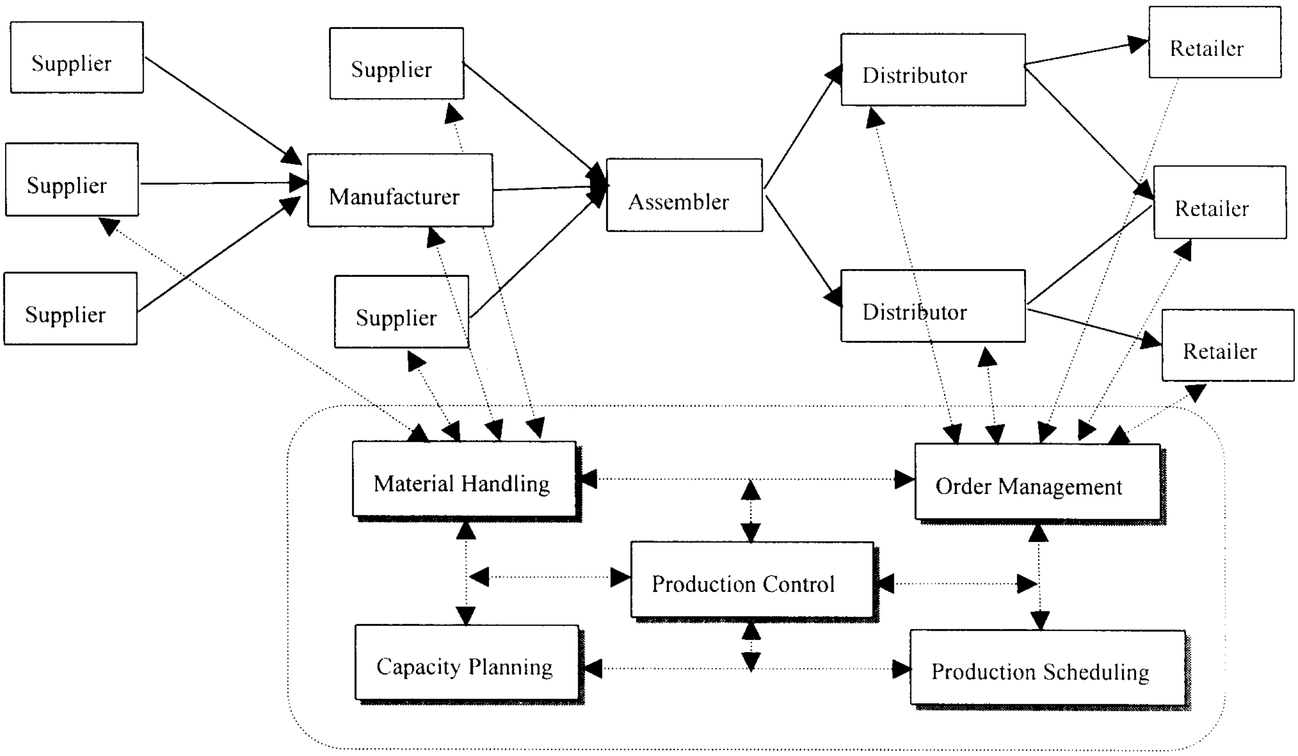


Figure 4. Order fulfillment process (OFP) within a supply chain network (Lin et al., 1998). Solid lines denote material flows while dotted arrows indicate information flows.

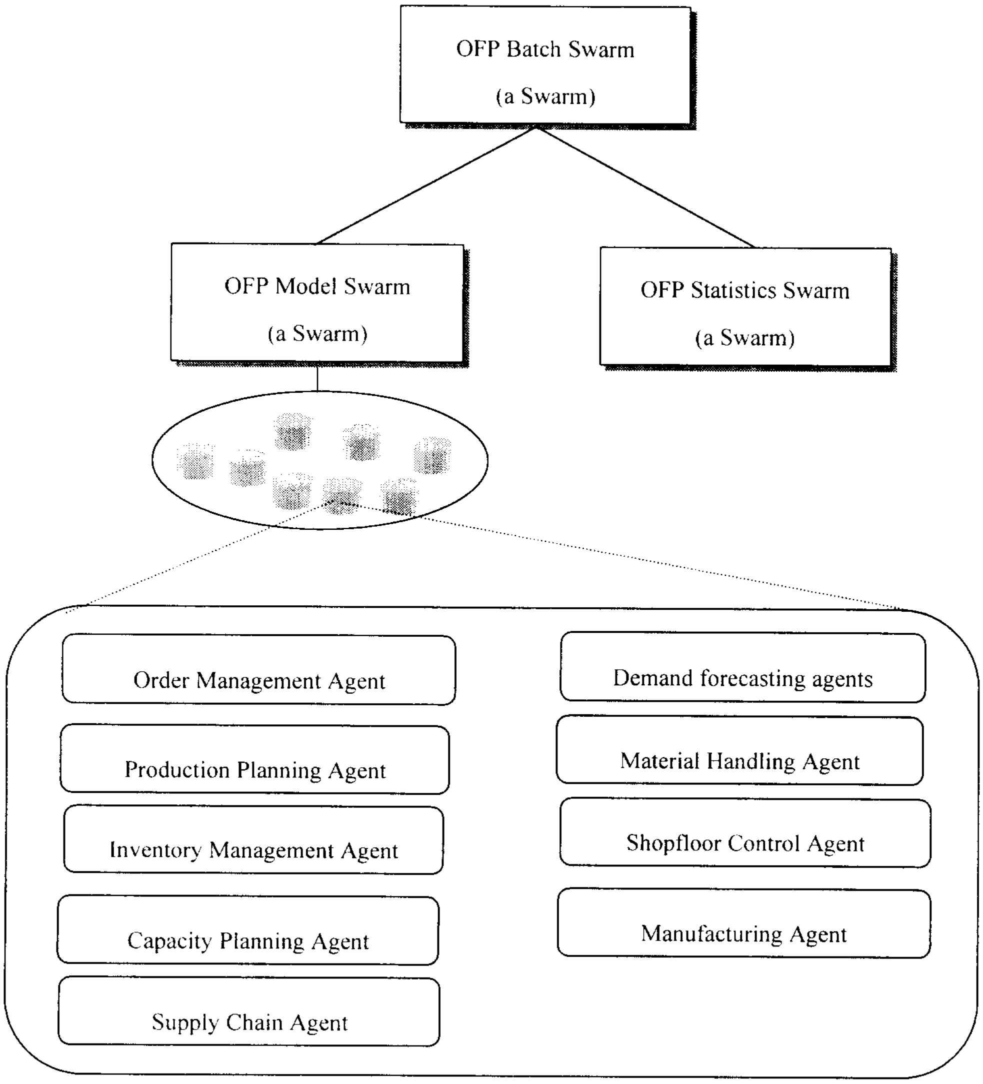


Figure 3. The implementation of Swarm in modeling a supply chain network (Lin et al., 1998).

- Consider the vector of feature weights at time t during training.

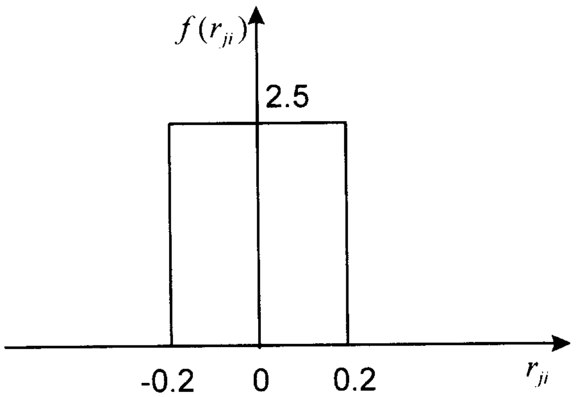
$$\mathbf{q}_i(t) = \begin{bmatrix} q_{1i}(t) \\ \vdots \\ q_{mi}(t) \end{bmatrix}$$

- At time $t + 1$, the subsequent value for each weight is

$$q_{j,i}(t+1) \equiv (1 + r_{ji}) q_{ji}(t) \text{ for } j = 1,\ldots,m$$

Here the scaling factor r_{ji} is random number uniformly distributed between $[-a, +a]$.

- In this study, $a \equiv 0.2$. The corresponding uniform density function is shown below.



- If the change in error E_i is favorable (that is, $E_{i+1} < E_i$), then $\mathbf{q}_{i+1}$ is adopted as the new weight. Otherwise, another vector $\mathbf{r}_i$ of scaling factors is selected.

Figure 4. Illustration of a multistrategy procedure for optimizing case based reasoning through directed random search.