VERSITA

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **12**, No 1

Sofia • 2012

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2012-0004

Insurance Dynamics – A Data Mining Approach for Customer Retention in Health Care Insurance Industry

V. Sree Hari Rao, Murthy V. Jonnalagedda***

** Professor of Mathematics, JNTUH, Hyderabad*
*** Professor of CSE, KNTUK, Kakinada*
Emails: vshrao@yahoo.com mjonnalagedda@gmail.com

***Abstract*:** *Extraction of customer behavioral patterns is a complex task and widely studied for various industrial applications under different heading viz., customer retention management, business intelligence and data mining. In this paper, authors experimented to extract the behavioral patterns for customer retention in Health care insurance. Initially, the customers are classified into three general categories – stable, unstable and oscillatory. To extract the patterns the concept of Novel index tree (a variant of K-d tree) clubbed with K-Nearest Neighbor algorithm is proposed for efficient classification of data, as well as outliers and the concept of insurance dynamics is proposed for analyzing customer behavioral patterns.*

Keywords: *Insurance dynamics, K-d tree, KNN, Novel Index tree, Customer retention.*

1. Introduction

Majority of the real world problems faced by industry come from the Insurance, Banking, and Telecom sectors. For an insurance company, a customer can be defined as a person, firm, or any other organization having one or more insurance policies in the same or different lines of insurance cover. The insurance provider collects a great deal of statistical information about the customers, such as features of insurance coverage, new policies, claims, renewals, cancellations, and so on. This information will be utilized by the provider to develop marketing strategies of the company. Usually insurance companies will be largely interested in those

customers who pay expensive premia and have only less number of claims. The main issue is to acquire, develop and retain core customer relationships. The following are the main objectives of any insurance company:

- running the company without violating any regulations;
- achieving higher profit margins on existing customers;
- retaining profitable customers for longer periods of time;
- fulfillment of commitments to the members and their dependents in terms of assuring quality service.

In the process of fulfilling these objectives, the insurance companies consider such issues, as loyalty of customers, fraud detection, reduction of high-risk members, and data acquisition. The present case study essentially deals with the issue of loyalty of customers. There are various definitions that highlight the loyalty of customers and we refer the readers to [7, 8, 13, 19, 20, 21, 23, 27].

It is observed that the behavior pattern of customers is a time dependent dynamic process. This necessitates one to deal with the question of retention of customers from a dynamical system point of view. There are several types of insurance policies that deal with automobiles, health, life, property, and so on. We limit the scope of this article to address the question of retention of customers in the health insurance sector.

The present paper is organized as follows:

Section 2 deals with the dynamical activity of the insurance industry, classifying the customers based on their behavior. The loyalty of customers varies with time depending on several factors. Hence, the concept of insurance dynamics is proposed.

Section 3 introduces a data structure called Novel Index Tree and a new search algorithm. They are variants of the K-d tree and the K-Nearest Neighborhood search algorithm, respectively. The data need to be classified for extracting useful patterns that are appropriate in knowledge extraction. A combination of the K-NN and the K-d tree is most convenient for classification of large data sets. Still, some limitations, such as bad spread data, computational overheads and outlier classification need to be addressed. This is accounted by Novel Index Tree and a new search algorithm.

Section 4 explains the proposed algorithm, its experimental results of classification and the interpretation of the classification.

Section 5 determines the distribution of the population of size 146 783, among stable, unstable, and oscillatory classes, using the proposed algorithm. Conclusions and discussion of the results are also presented in this section.

2. Insurance dynamics

Based on the behavioral patterns the customers can be classified into three general categories – stable, oscillatory, and unstable.

- **Stable.** The members of this category are those who join a particular insurance provider and continue staying with the same provider. This behavior needs not necessarily be because of complete satisfaction with the plan (usually

such people are averse to change). These members usually tend to add stability to the provider.

- **Oscillatory.** The members of this category are those who would frequently change providers and in the process, would revisit the provider with whom they stayed for a while in the past. This behavior is an indicator of their indecisiveness and indetermination. This induces a sort of oscillatory tendency in the form of back and forth movement among providers/plans.
- **Unstable.** The members of this category are those who are determined in their decision to leave the provider, with no possibility of returning. Clearly from the point of view of the provider, this category exhibits instability characteristics.

These considerations would make one believe that the dynamical system approach is appropriate for exploration in the insurance industry.

These notions will be discussed in the next section. We observe that the behaviour of customers will be time dependent. Hence, we view this as a dynamical process and designate this as Insurance Dynamics.

We now formulate the following definitions:

A. Stable class

We say that a customer X is stable if $X(0) \in P \Rightarrow X(n) \in P$ for all $n \geq 0$, where 0 denotes the enrollment of X into P, initially P denotes population that belongs to a plan or a provider in general.

This means that a customer who enrolled in a plan/provider initially, continues to be with the same plan/provider for all future time.

B. Oscillatory class

Let N be the set of natural numbers. At instant i, let sgn(X) denotes the sign of the variable $X(i)$.

We say that X is an oscillatory customer if there exist i and j in N, such that $\text{sgn}(X(i)X(j)) < 0$ for all $j > i \geq 0$.

Here $X(i)$ is assigned a positive sign if $X(i)$ belongs to P for that i, and $X(j)$ is assigned a negative sign if $X(j)$ does not belong to P.

The members of the oscillatory class are further divided into two categories, viz., a) weakly oscillatory class and b) strongly (chaotic) oscillatory class.

For a member of oscillatory class, we define the oscillatory period as the number of discrete time units elapsed in respect of an individual enrolled with P for leaving and returning to P. That is, the oscillatory period is equal to the numerical difference between the instances that correspond to leaving and returning. Obviously a member of an oscillatory class may have varying oscillatory periods. If these periods are in the increasing order for an individual we term the oscillation of that individual as strongly oscillating or chaotic. If these periods are either constant or in the decreasing order then we term the oscillation of that individual as weakly oscillatory. It is interesting to note that chaotic oscillations may exhibit a tendency of instability whereas the weak oscillation may tend to stabilize eventually.

C. Unstable class

A customer who is not stable or oscillatory in the sense of the above definitions is termed as an unstable customer.

3. A novel classification technique

The interest in utilizing data mining technique has been continually increasing in recent among actuaries as is apparent from the large literature on this subject [6, 9-12, 15, 21, 23-26, 28] and the references therein. Data mining is based on modern, computer – intensive methods that are reliable and fast. The effectiveness of data mining techniques lies in the following:

Data mining approach overcomes the shortcomings of traditional methods which rest on the assumption that the data are distributed normally or according to certain distribution law such as binomial, Poisson or Gamma. This assumption is not always correct. Also, the data mining methods rely on the intense use of computing power which renders the analysis less time consuming. On the other hand classical methods applied to large data sets takes longer time. Further, these methods are capable of handling categorical variables with a large number of categories such as claimant's postal code, organization, occupation etc. Traditional methods face problems to deal with the categorical variables and will be left out or need to be grouped by hand before inclusion. Besides the data mining methods such as decision trees are capable of handling noisy or incomplete data which sometimes creates problems for the application of linear models. The data mining methods in insurance have been applied to analyze large data sets. Such methods are employed for risk prediction / assessment, premium setting, fraud detection, health costs prediction, member retention and acquisition strategies, treatment management optimization and investments management optimization (see [9, 28]). The differences between data mining and on-line analytical processing (OLAP) methodologies have been well explained in [11]. For the use of other techniques such as classification and regression trees (CART) and multivariate adaptive regression splines (MARS), we refer the readers to [6, 10, 12, 17, 21, 22, 24-26].

The objectives of an insurance company are running the company without violating any regulations, achieving profit margins on new customers within short time, achieving higher profit margins on existing customers, retaining profitable customers for a longer period of time, and fulfillment of commitments to the members and their dependents in terms of assuring quality service. Classification methods are used to identify type of a customer based on his policy / insurance characteristics. In the current application of customer retention we consider two major factors to classify the population as stable, unstable and oscillatory. They are the current continuous stay and the oscillatory indicator. The unstable customers will join the policy once and they discontinue and will not come back again as they were not happy with the policy.

The stable customers join the policy and continue to stay with the company for longer period.

The oscillatory customers join the policy, stay for some time and switch over to another policy as some features are attractive and come back to the original plan as there is some change in the policy. They dislike some features in the policy and are not fully satisfied with any policy, and keep switching the providers.

A customer would have enrolled into a plan some time back and continuously existing with the same plan till date. This duration of stay can be defined as current continuous stay.

The oscillatory indicator shows how many times the customer has switched his policy and has come back to the current provider.

These two parameters play a dominant role in classifying the customers.

Among the classification methods K-nearest neighborhood method (K-NN) is very popular. Also for an efficient-classification and scalability KNN is combined with K-d tree [1, 5, 19].

In conventional K-d-tree [19], hyper rectangles which intersect a hyper circle around the point of interest are used for nearest neighbor search. This becomes the major source for computational burden, we propose to consider those hyper rectangles that intersect the circle topologically. Further, by storing certain parameters such as minimum and maximum limits along each dimension, sum and the number of elements in intermediate (non-leaf) nodes, a great deal of calculation overhead is reduced. The outlier classification problem is addressed by class limits. Hence, we propose to construct a new tree called Novel index tree and a new search algorithm.

In a K-d tree, search cost rises exponentially with “k” (No of keys) [2-5]. Many methods are available in literature to reduce search time [16]. L e e and H y o u n g[16] proposed to divide d-dimensional data space into 2-dimensional spherical pyramids. In [14], an index structure for higher dimensional nearest neighbor queries has been proposed. We study the same problem from a different perspective and propose that in multidimensional space the hyper-cuboid that represents the search space can be approximated by a hyper sphere, which avoids a great deal of computational overhead.

As a case study, this algorithm is applied to customer retention problem in health care insurance. The execution time of the proposed algorithm is compared with SALFORD Corporation’s CART 5.0 algorithm for a test data of about 150 000 records. Also, how this algorithm may be extended to higher dimensional spaces is explained.

A non-recursive divide and conquer procedure is adopted to build the tree. The arithmetic mean of the elements on a particular dimension is computed. The elements with value less than or equal to the mean go to the left sub tree and the rest go to the right sub tree. The process repeats on other dimensions also in subsequent levels, until leaf nodes are obtained with single element. The training data and the test data will be stored in flat files. While building the tree, we store the minimum and maximum limits of elements on each dimension in the intermediate node for reducing the calculation overhead.

Another important consideration for reducing the execution time is to consider those hyper rectangles, which intersect with the circle topologically for nearest

neighborhood search. While classifying the elements, the boundaries of each class can be computed, this brings efficiency in outlier classification.

D. Intersection of a rectangle and a circle

In order to determine whether a rectangle and a circle intersect, the following cases may be considered.

Case I. The center L of the circle lies inside the rectangle. Let $O(x_m, y_m)$ is the center of the rectangle. In this case, clearly the circle and the rectangle intersect each other.

Let L_1, L_2, L_3 and L_4 be the four sides of the rectangle $P_1(x_1, y_1)$, $P_2(x_2, y_1)$, $P_3(x_2, y_2)$, $P_4(x_1, y_2)$ and let d_1, d_2, d_3 and d_4 be the perpendicular distances from $L(x_{val}, y_{val})$ to L_1, L_2, L_3 and L_4 respectively as shown in Fig. 1.

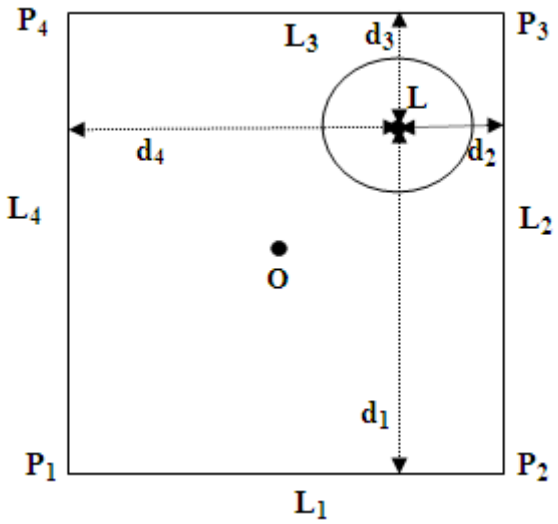

Fig. 1. Center of the circle lies inside the rectangle

If a point ***L*** lies inside the rectangle, the following conditions hold:

$$d_1 + d_3 = y_2 - y_1 \quad and \quad d_2 + d_4 = x_2 - x_1.$$

Case II. The center of the circle lies outside the rectangle.

If a point L lies outside the rectangle, then $d_1 + d_3 = y_2 - y_1$ and $d_2 + d_4 = x_2 - x_1$ cannot hold simultaneously. If L lies outside, then one of the conditions will fail depending upon the position of L.

The following situations arise in this case**:**

1. The circle intersects the rectangle (see Fig. 2).
2. The circle does not intersect the rectangle and lies outside the rectangle (see Fig. 3).

Let us derive the condition for checking whether the circle and the rectangle intersect or not. The line joining O and L will intersect any two sides of the rectangle and extended lines of the other two sides at utmost four points.

Let the co-ordinates of the nearest corner point of the rectangle from Centre of

circle L be (x_2, y_2) where the two perpendicular sides will intersect. The line joining O and L will intersect L_2 at $M_1(x_2, y_{int1})$ and extended line of L_3 at M_2 (x_{int2}, y_2). Out of M_1 and M_2 one point will be internal (M_1) and other point (M_2) will be external. Let the co-ordinates of farther points that lie on sides parallel and perpendicular to X-axis be (x_1, y_2) and (x_2, y_1) respectively.

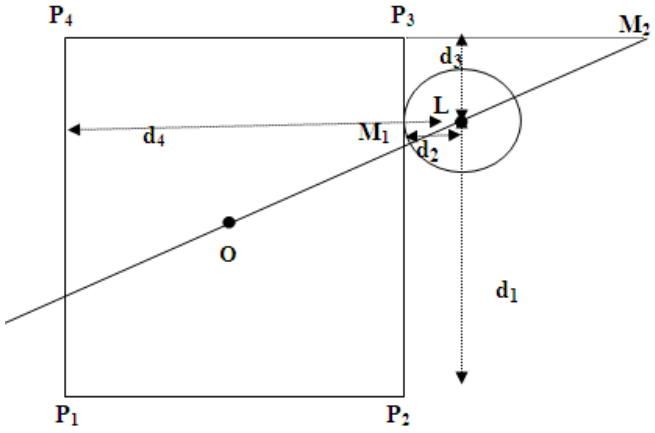

Fig. 2. The circle intersects the rectangle

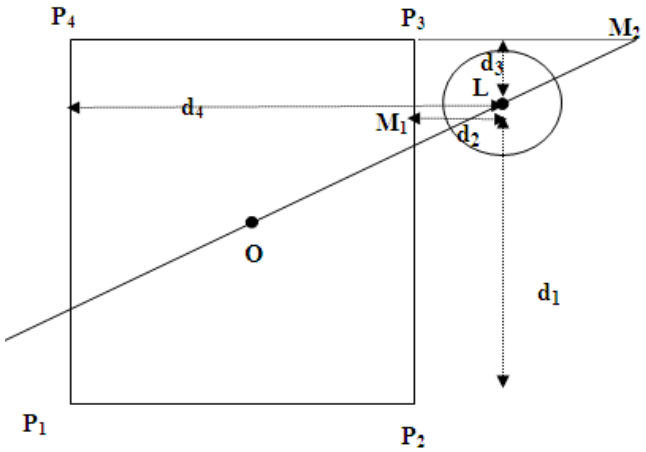

Fig. 3. The circle does not intersect the rectangle

The co-ordinates of O are (x_m, y_m) and the co-ordinates of L are (x_{val}, y_{val}). Then the equation of OL will be

$$(y - y_m) = m(x - x_m),$$

in which,

$$m = (y_{val} - y_m)/(x_{val} - x_m).$$

Thus $y - y_m = m(x - x_m)$ and hence

(3.1) $$y = mx + (y_m - mx_m).$$

The equations of lines passing through (x_2, y_2) are

(3.2) $X = x_2$ (parallel to Y-axis),

(3.3) $Y = y_2$ (parallel to X-axis).

Solving (3.1) and (3.2) we get the co-ordinates of the point M_1 and accordingly $y_{int1} = mx_2 + (y_m - mx_m)$ and $M_1 = (x_2, y_{int1})$, $d_1 = |y_1 - y_{int1}|$ (perpendicular distance to L_1 from M_1 (perpendicular distance to the far point is to be considered).

Also, solving (3.1) and (3.3) we get the co-ordinates of point M_2 as $x_{int2} = (y_2 + mx_m - y_m)/m$ and consequently $M_2 = (x_{int2}, y_2)$, $d_2 = |x_1 - x_{int2}|$ (perpendicular distance to the far point).

Now, either M_1 or M_2 should be an internal point to the rectangle. Let the internal point be M. If the perpendicular distance from the intersection point to the farther point is greater than the length of that side, it is considered to be an external point otherwise it is an internal point to the rectangle.

Here $|y_1 - y_{int1}| < |y_1 - y_2|$ hence M_1 is an internal point.

If $OM + r \geq OL$, then the circle and the rectangle intersect, otherwise they do not intersect.

The following special cases arise in **Case II:**

1. The rectangle transforming into a line: In this case the perpendicular distance from the line to the point should be less than **r**. This condition arises when $x_1 = x_2$ or $y_1 = y_2$. The distance $d = |x_{val} - x_1|$ or $|y_{val} - y_1|$ depending on the condition.

2. The line joining the centers of rectangle and circle is parallel to either X-axis or Y-axis. If $(x_{val} = x_m)$ the line joining the centers is parallel to Y-axis.

Set $l = y_m - y_1$, d = distance between (x_{val}, y_{val}) and (x_m, y_m)

If $(y_{val} = y_m)$ the line joining the centers is parallel to X-axis. Then $l = x_m - x_1$

If $d \leq (l+r)$ then the rectangle and circle intersect

3. One of the sides of a rectangle is tangential to the circle and this condition is true if $|x_2 - x_{val}| = r$ or $|y_2 - y_{val}| = r$.

4. The circle may be tangential to the vertices of the rectangle and the distance between (x_{val}, y_{val}) and (x_n, y_n) will be equal to 'r'.

If any points are obtained in the nearest neighbor the class of the majority points is assigned to the point (x_{val}, y_{val}) otherwise the record is written to the outlier file.

Unclassified samples are:

Generally the outlier classification process is as follows:

Consider one outlier point. Compute its distance with all the points in a particular class and find out the minimum distance point in the class. Similarly compute the minimum distance with each of the classes. Find out the minimum among the minimum class distances. The point may be classified as belonging to the minimum distance class.

Now we present the following Algorithm:

Step 1. Read the data from flat files and load it into an array.

Step 2. Compute the arithmetic mean of the array elements.

Step 3. Divide the array into sub arrays, left and right array.

Step 4. Push the right array into a stack.

Step 5. Consider the left array as array and repeat Steps from 2 up to 4 till a single element array is generated.

Step 6. Pop the right array from the stack and repeat Steps from 2 up to 5 till the stack becomes empty.

Step 7. Read the test data points.

Step 8. For every point in the test data classify the points and find out the class limits. If the point is not classified store it into outlier file.

Step 9. Classify the outlier using the class limits.

4. Experiments

We have experimented on a data set of 146 713 records and they were classified with 100% accuracy. Consider a training data set of consisting of about 14 000 records. If we look at the distribution of the population 41.46% of them belong to the stable class, 23.79% of them belong to unstable class and 34.75% belong to Oscillatory class. Among the oscillatory customers around 50.30% are weakly oscillatory.

Table 1. Distribution of the population among various classes

Type of Class	**No of persons**	**Percentage**
Stable	60832	41.46%
Unstable	50986	34.75%
Oscillatory	34895	23.79%
Strongly	17342	11.82%
Weakly	17553	11.97%

When implemented our algorithm took about 6.5 seconds to classify 146 713 records. When the same was implemented on **SALFORD** Corporation's **CART 5.0** package it took around 15 seconds to classify the same data.

Now looking from the business perspective the company should target the oscillatory class customers for improving the business. The common properties of oscillatory class people, that influence their retention should be analyzed and then the reason for their oscillatory behavior can be found.

Stable class customers can be given same incentives to make them more loyal. This kind of analysis of customer behavior patterns will help business to grow.

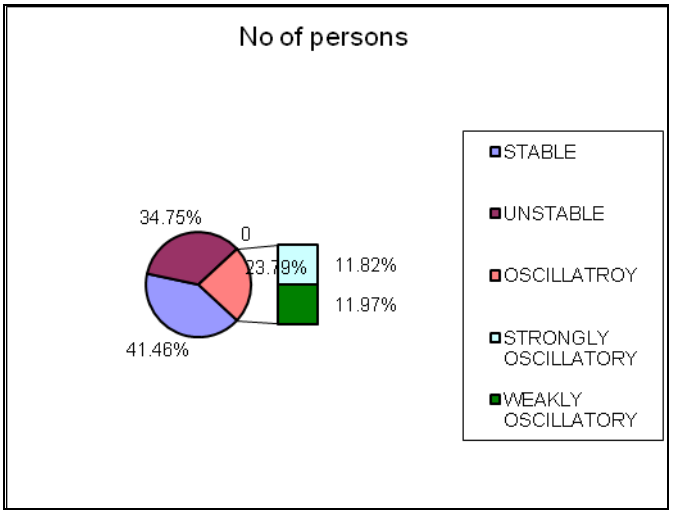

Fig. 4. Population distribution on customer category

5. Conclusions and discussion

In this paper the authors considered the issue of customer retention in health care insurance industry. Observing that the behavior patterns of customer is a time dependent dynamical process; the authors proposed the notion of insurance dynamics as a fundamental step to address the question of customer retention in health care insurance industry. The total population of the insurance customers is divided into three mutually exclusive classes such as stable, unstable and oscillatory classes. The oscillatory class is further divided into weakly oscillatory and strongly oscillatory (chaotic) classes. The usefulness of this dynamics has been tested on a simulated data size of about 150 000 records.

The proposed methodology of introducing topological intersection, storing min-max limits in the intermediate nodes and class limits concept has rendered a considerable reduction in the execution time for low dimensional data.

Moreover, from the proposed methodology, the following tasks appear to be the next step in terms of retention of customers:

- This algorithm can be extended to higher dimensional data by approximating the hyper-cuboids with hyper-spheres. A detailed financial analysis is required to estimate the break-even costs for deciding the premium. Then various control mechanisms can be introduced to stabilize the system.
- A more detailed analysis of the effect of “events in the customers life” (marriage, divorce, death of spouse, change of job, children, change of income, retirement, voluntary retirement, migration to another country etc.)
- A time continuous approach to score loyalty which would introduce new dynamics into the system, also by considering time between events occurring in the life of a customer.
- An application of survival analysis techniques, by considering the professional and occupational hazards that induce health problems (for example

people working in asbestos industry are subject to cancer risks, etc.), which finally effects the loyalty aspects of the customer.

References

1. Bently, J. Multidimensional Binary Search Trees Used for Associative searching. – Comm. ACM, Vol. **18**, 1975, No 9, 509-517.
2. Berchtold, S., C. Bohm, H-P. Kriegel. The X-Tree Indexing Structure for High-Dimensional Data. – In: Proc. 22nd Int. Conf. Very Large Database, September 1996, 28-39.
3. Berchtold, S., C. Bohm, H.-P. Kriegel. A Cost Model For Nearest Neighbor Search in High-Dimensional Data Space. – In: Proc. ACM PODS Symp. Principles of Database Systems, 1997.
4. Berchtold, S., C. Bohm, H.-P. Kriegel. The Pyramid-Technique: Towards Breaking the Course of Dimensionality. – In: Proc. ACM SIGMOD Int. Conference. Management of Data 1998.
5. Beyer, K., J. Goldstein, R. Ramakrishnan, U. Shaft. When Is "Nearest Neighbor" Meaningful?– In: Proc. Seventh Int. Conf. Database Theory, January1999, 217-235.
6. Breiman, L., J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees. Wadsworth, Pacific Grove, California, 1984.
7. Brown, G. H. Brand Loyalty – Fact or Fiction? – Advertising Age, Vol. **9**, 1952, 53-55.
8. Fournier, S., J. L. Yao. Reviving Brand Loyalty: A Reconceptualization Within the Framework of Customer-Brand Relationships. – International Journal of Research in Marketing, Vol. **14**, 1997, No 5, 451-472.
9. Francis, L. Neural Networks Demystified. Casualty Actuarial Society Forum, 2001, 252-319.
10. Haberman, S., A. E. Renshaw. Actuarial Applications of Generalized Linear Models. – In: D. J. Hand, S. D. Jacka, Eds. London, Statistics in Finance, Arnold, 1998.
11. Han, J., M. Camber. Data Mining: Concepts and Techniques. New Delhi, India, Morgan Kaufmann Publishers, An Imprint Elsevier, 2001.
12. Hastie, T., R. Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. – New York, Springer-Verlag, 2001.
13. Jacoby, R. Chesnut. Brand Loyalty: Measurement and Management. New York, Wiley, 1978.
14. Katayama, N., S. Satoh. The SR-Tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. – In: Proc. ACM SIGMOD Int. Conference. Management of Data, May 1997, 517-542.
15. Kolyshkina, I., D. Steinberg, N. S. Cardell. Using Datamining for Modeling Insurance Risk and Comparison of Datamining and Linear Modeling Approaches. Chapter 14. – In: Intelligent and Computational Techniques in Insurance – Theory and Applications. A. F. Shapiro, L. C. Jain, Eds. Vol. **6**. World Scientific Publications, 2003.
16. Dong-Ho, Lee, Kim Hyoung-Joo. An Efficient Technique for Nearest Neighbour Query Processing on the SPY – TEC. – IEEE Transactions on Knowledge and Data Engineering, Vol. **15**, 2003, No 6, 1472-1486.
17. Lewis, P. A. W., J. Stevens, B. K. Ray. Modeling Time Series Using Multivariate Adaptive Regression Splines (MARS). – In: A. Weigend, N. Gershenfeld, Eds. Time Series Prediction: Forecasting the Future and Understanding the Past, Santa Fe Institute: Assison-Wesley, 1993, 297-318.
18. Lin, K. L., H. V. Jagadish, C. Faloutsos. The TV-Tree: An Index Structure for High-Dimensional Data. – The Very Large Data Bases J., Vol. **3**, 1994, No 4, 517-542.
19. Moore, A. Efficient Memory – Based Learning for Robot Control. Ph. D Thesis, University of Cambridge, 1991.
20. Mowen, J. C. Customer Behaviour. New York, Prentice Hall, 1995.
21. McCullagh, J. A. Nelder. Generalized Linear Models. 2nd Ed. London, Chapman and Hall, 1989.

22. Salford Systems, Multivariate Adaptive Regression Splines (MARS), 2002.
http://www.salfordsystems.com
23. S h a p i r o, A. F., L. C. J a i n. Intelligent and Other Computational Techniques in Insurance – Theory and Applications. – Series on Innovative Intelligence, Vol. **6**, World Scientific Publications, Singapore, 2003.
24. S m y t h, G. Generalized Linear Modeling. 2002.
http://www.statsci.org/glm/index.html
25. S t e i n b e r g, D., N. S. C a r d e l l. Improving Data Mining with New Hybrid Methods. Boston, MA, DCI Database and Client Server World, 1998.
26. S t e i n b e r g, D., N. S. C a r d e l l. The Hybrid CART – Logit Model in Classification and Datamining. – In: Eight Annual Advanced Research Techniques Forum, American Marketing Association, Keystone, Co., 1998.
27. U n c l e s, M., G. L a u r e n t. Editorial. – International Journal of Research in Marketing, Vol. **14**, 1997, No 5, 399-404.
28. Work Cover NSW News, Technology Catches Insurance Fraud, 2001.
http://www.workcover.nsw.gov.au/pdf/wca46.pdf