2011 3rd International Conference on Information and Financial Engineering
IPEDR vol.12 (2011) © (2011) IACSIT Press, Singapore

A Risk Recognition Model in Finance Based on Graphical Knowledge Discovery

Long Wu [1], Mingfei Xie [2]

[1] Shenyang Normal University

[2] Shenyang Normal University

Abstract. In this paper, the Lundberg risk model which perturbed by diffusion is extended to incorporate the jumps of surplus investment returned. Under the supposition that the jump of surplus investment return follows a multiplex Poisson process with Laplace disseminated jump sizes, through the factorization technique instead of the integro-differential equation approach, we obtain the explicit closed form expression of the resulting it expected discounted penalty function. In particular, when the claim dissemination is of Phase-type, the expression of the EDP function is simplified even further as a compact matrix-type form. Ultimately, the financial applications include pricing barrier option and everdurimg American put option and determining the optimal capital structure of a firm with endogenous default.

Keywords: risk, finance, graph, pattern recognition.

1. Introduction

The Lundberg risk model has been researched widely in ruin theory and references therein. Dufresne and Gerber thrust the Lundberg risk model by adding a radiating process with sustaining volatility. The accessional radiating term may be commentated as the future uncertainty of aggregate claims, the future uncertainty of premium incomes, or the undulation of investment of surplus. This new model is usually called the perturbed compound Poisson risk model in the context of ruin theory.

As a result, so many quantities of interest are written in a closed form. Further more, when the demand size is of Phase-type, Ren applies the results of Lin and Willmot to acquire matrix representations of the expected time of ruin and the moments of the discounted deficit at ruin. They also focus on Phase-type distributed claims, use an ordinary differential equation approach to solve this IDE and consequently obtain a simple explicit solution. Ren's results are further gathered up by Chi in which the EDP function for a risk model with stochastic volatility is considered. With a martingale approach, Chen derives an IDE for the EDP function. Chen consider a risk model that allows hyper-exponential upward jumps[1]. As an application, they work out an example in Leland's structural model of corporate default.

2. The Wiener Factorization

The ecumenic idea of Wilcoxon rank-sum test is that, instead of using the primordial observed data, we can list the data in the raising order, and assign each data item a rank, which is the place of the item in the sorted list. Then, the ranks are used in the analysis. Using the ranks instead of the original observed data makes the rank sum test much less sensitive to noises than the classical parametric tests. Generally, noise will change the tstatistic value greatly, but change the ranks little. Since a gene expression dataset often has much noise, it is more suitable for applying the Wilcoxon rank-sum test on informative gene selection[2].

One character of microarray data is that the amount of tumor samples collected tends to be much smaller than the amount of genes. The former tends to be on the order of tens or hundreds, whereas the latter usually contains thousands of genes on each chip. In statistical term, the number of predictor variables exceeds the number of samples by a considerable margin[3]. Obviously, this high dimensionality will unavoidably worsen the generalization performance of machine learning methods.[1]

Long Wu. Tel.:15040316135.
E-mail address: *wulongwin@hotmail.com.*

3. Analysis of the Graphical Function

3.1 Gene Expression Profiles

Simultaneously, a great quantity of genes are superfluous and do not contribute to cancer diagnosis. As a dimensionality reduction technique, feature extraction plays a significant role in tumor classification, because there are too many superfluous genes and too much noise existing in gene expression profiles. In fact, for pattern (tumor) classification, feature extraction can be viewed as a searching process for the best feature transformations which retain as much classification information as possible. Generally speaking, feature extraction can yield four advantages: Improving classification performance; Reducing the time complexity of classification model; Visualizing classification results when mapping the data into two- or three-dimensional space; Gaining biological insights[4].

$$\sigma^2 ab + cb - \lambda_2 \mathbb{E}\left[e^{-aY_1} \sin(bY_1)\right] + \frac{\lambda_1 \theta}{2}\left[\frac{b}{(\theta - a)^2 + b^2} - \frac{b}{(\theta + a)^2 + b^2}\right] = 0. \tag{1}$$

Usually, most of the existing methods adopt two-stage methods to select informative genes or extract features.

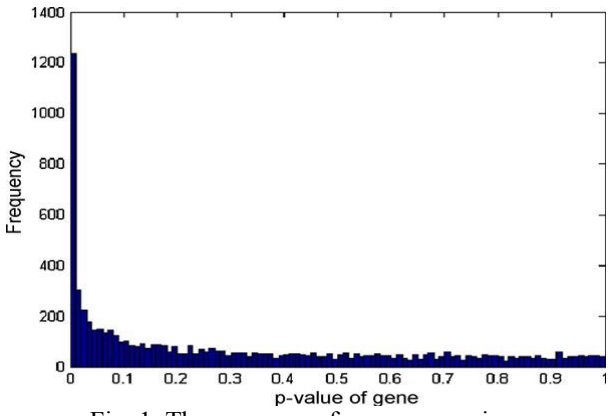

Fig. 1. The accuracy of gene expression.

3.2 Discrete Cosine Transform

One of the common ways and means to research the EDP function is to express the function as a solution of an integral equation or an integro-differential equation. Gerber and Landry show that the EDP function contents a defective renewal equation and provide a natural probabilistic interpretation of the equation. Lin and Willmot study the defective renewal equation of this kind and obtain an explicit closed-form solution in terms of a compound geometric distribution function[5].

$$\phi(\xi) = \lambda_1 \theta \left(\frac{1}{\theta + \xi} + \frac{1}{\theta - \xi}\right) \Big/ 2 - \lambda_1 = \lambda_1 \frac{\xi^2}{\theta^2 - \xi^2}. \tag{2}$$

An overview of the characteristics of all the datasets can be found in Fig. 2. The leukemia dataset consists of 72 bone marrow or peripheral blood samples including 47 acute lymphoblastic leukemia samples and 25 acute myeloid leukemia samples. The colon tumor dataset consists of 62 samples of colon epithelial cells including 40 colon cancer samples and 22 normal samples. The hepatocellular carcinoma data consists of 60 hepatocellular carcinoma tissues of which 20 suffer from early intrahepatic recurrence and 40 do not.

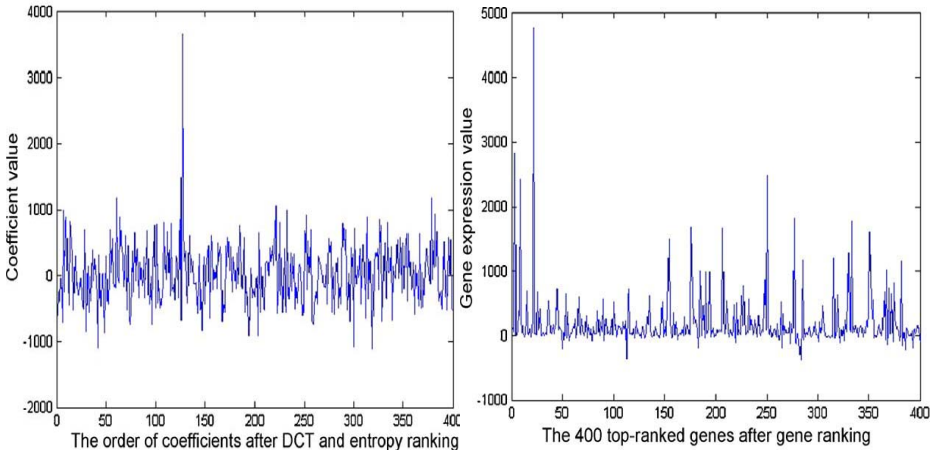

Fig. 2. The accuracy as a function of discrete cosine transform.

4. Experiment

4.1 Analysis of the Free Parameters

Those differentially expressed genes by rank sum test were first selected. To denoise those informative genes we further adopted DCT to transform the expression values of those selected genes into the coefficients so as to compact the energy[6]. However, it is possible that the coefficients with high energy contribute little to classification, so an entropy ranking based method was adopted to sort the transformed coefficients so that a set of coefficients with high classification ability can be selected. To drastically reduce the dimensionality of the selected coefficients, PCA was adopted to extract PCs which were used as the inputs of the classifier. The formula is:

$$\phi^{-}(s) \triangleq -\frac{\delta}{\phi(s)\phi^{+}(s)} = -\frac{\delta\theta}{\xi_1\xi_2} \times \frac{(\xi_1 - s)(\xi_2 - s)}{(\theta - s)\phi(s)}, \quad \mathscr{R}s = 0, \tag{3}$$

In my previous work, we employed independent component analysis to model the gene expression data, then applied optimal scoring algorithm to classify them. For the sake of comparison, penalized ridge regression, penalized principal component regression, and PAM were also used to do the same tumor classification experiment. In my previous work, we proposed a multi-step dimensionality reduction method proved to be successful and effective to classify the tumor dataset[7]. The experimental results showed that the obtained classification performances were very steady, as evaluated by SVM and K-NN classifier.

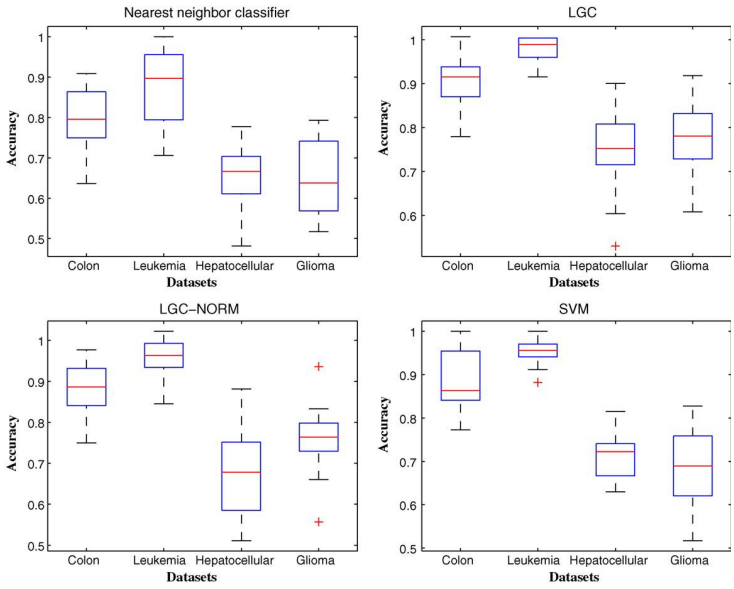

Fig. 3. The accuracy as a function of the free parameter.

4.2 The Relationship of The Accuracy

The parameter s is in Step 1 of the algorithm LGC in Section 2.5. Fig. 4shows that other than hepatocellular carcinoma data, the bests for the other three datasets is 0.7.The bests for hepatocellular carcinoma data is about 1.2. The overall tendency of the accuracy decreases with the increase of s, but it is not monotonically decreasing. It presents a sawtooth-like decrease. Fig. 5 shows that when LGC-NORMis performed, there is no better method to choose the bests. The formula of tendency is:

$$\begin{aligned} \frac{\phi(s)}{(s-\xi_1)(s-\xi_2)} &= \frac{\frac{\phi(s)}{s-\xi_1} - \frac{\phi(\xi_2)}{\xi_2-\xi_1}}{s-\xi_2} \\ &= D + \lambda_2 \frac{\frac{\check{P}(s)-\check{P}(\xi_1)}{s-\xi_1} - \frac{\check{P}(\xi_2)-\check{P}(\xi_1)}{\xi_2-\xi_1}}{s-\xi_2} + \frac{\lambda_1\theta}{2} \\ &\times \left(\frac{1}{(\theta+\xi_1)(\theta+\xi_2)(\theta+s)} + \frac{1}{(\theta-s)(\theta-\xi_1)(\theta-\xi_2)} \right) \end{aligned} \tag{4}$$

The overall accuracy tendency for acute leukemia data increases with the increase of s while the overall tendency of the accuracy for the other three datasets has no general rule to follow[8]. Furthermore, it can be seen that the best s's for colon cancer data, high-grade glioma data, and hepatocellular carcinoma data within the interval [0.7, 1.6] are about 1.1, 0.8, and 0.7, respectively.

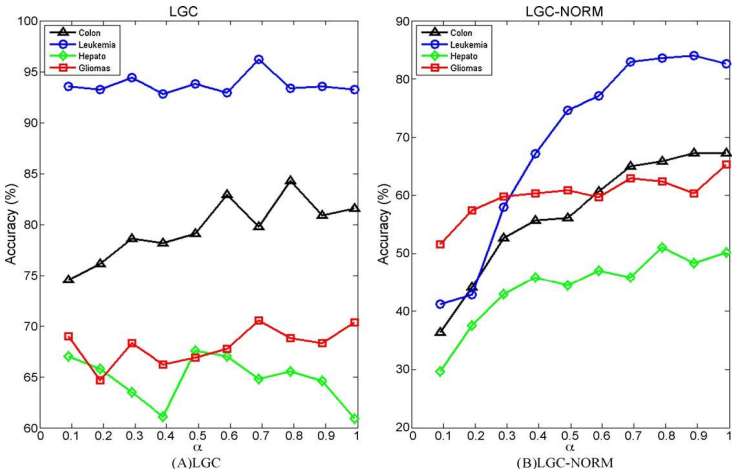

Fig. 4. The accuracy as a function of the high relational parameter.

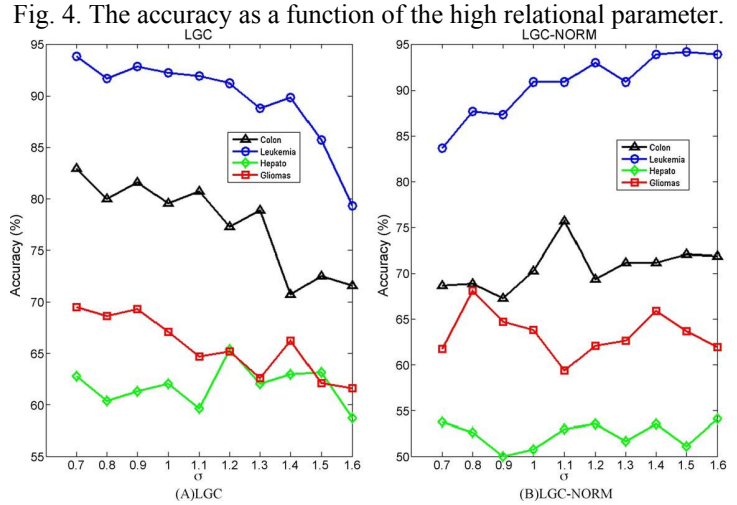

Fig. 5. The accuracy as a function of the low relational parameter.

4.3 The Relationship between The Accuracy and Label Data Size

Preprocessing of this dataset was done by setting threshold and log-transforming on the original data, which is similar to what was done in the original publication. The threshold technique is generally achieved by restricting gene expression levels to be larger than 20. In other words, the expression levels smaller than 20 will be set to 20. Regarding the log-transformation, the natural logarithm of the expression levels usually is taken. However, this pre-processing is not applied to the rest of the datasets[9].

Another experiment was conducted to show the relationship between the labeled data size and the accuracy. The results are depicted in Fig. 6. When the size of labeled data is 10%, we randomly selected 10% of the total dataset. It can be easily seen from Fig. 6 that the accuracy increases with the increase of labeled data size for both LGC and LGC-NORM.

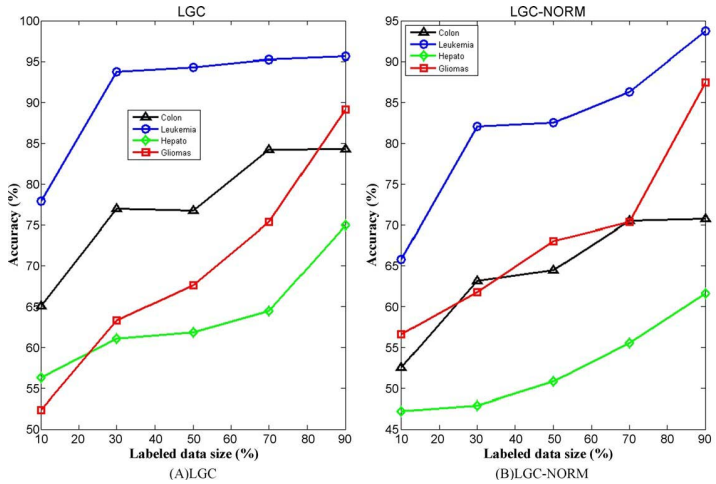

Fig. 6. The accuracy of the graphical model.

5. Conclusion

This paper extends my aforetime work which depicts the early efforts at tumor classification using gene expression data. Commonly, gene expression in cells is very complex, which leads to the tumor dataset obtained by microarray techniques including too many tumor-unrelated and redundant genes and too much noise.

In this paper, a semi-supervised graph-based measure, referred to as LGC, was used for classification, and its variant LGC-NORM was used too. The algorithm in this paper was tested on four well-known tumor datasets to assess the performance. Comparing the experimental results of the method with the ones of the other 13 methods, the results show that our method is indeed effective and efficient in predicting normal and tumor samples from four human tissues. In addition, SVM and K-NN classifier were used for comparison. It showed that the performance of LGC is the best while that of K-NN is the worst. The capability of SVM is competitive with that of LGC-NORM. Furthermore, these results hold under re-randomization of the samples. Another advantage for our approach is that the classification results can be visualized when extracting only three PCs with high accuracy.

6. Acknowledgements

This work was supported by Professor Liu.

7. References

[1] P. Barbe, Fougères, A.-L., and C. Genest. On the tail behavior of sums of dependent risks. *The Astin Bulletin*, 2006, 36: 361-373.

[2] J. Dhaene, R.J.A. Laeven, S. Vanduffel, G. Darkiewicz, and M.J. Goovaerts. Can a coherent risk measure be too subadditive? *Journal of Risk and Insurance*, 2008, 75: 365-386.

[3] S. Asmussen, F. Avram, and M.R. Pistorius. Russian and American put options under exponential phase-type Lévy models. *Stochastic Processes and their Applications*, 2004, 109: 79-111.

[4] J. Cai, and H.L. Yang, 2005. Ruin in the perturbed compound Poisson risk process under interest force. *Advances in Applied Probability*. 2005, 37: 819-835.

[5] Y.T. Chen, C.F. Lee, and Y.C. Sheu. An ODE approach for the expected discounted penalty at ruin in a jump-diffusion model. *Finance and Stochastics*. 2007, 11: 323-355.

[6] Y. Chi, and X.S. Lin. On threshold dividend strategy in a two-sided jump-diffusion risk model. *Working paper*. 2009.

[7] A.E. Kyprianou, and B.A. Surya. Principles of smooth and continuous fit in the determination of endogenous bankruptcy levels. *Finance and Stochastics*. 2007, 11: 131-152.

[8] J.D. Ren. The expected value of the time of ruin and the moments of the discounted deficit at ruin in the perturbed classical risk process. *Insurance: Mathematics and Economics*. 2005, 37: 505-521.

[9] E. Mordecki. Ruin probabilities for Lévy processes with mixed-exponential negative jumps. *Theory of Probability and its Applications*. 2004, 48: 170-176.