Evaluation of Predictive Learners for Cancer Incidence and Mortality

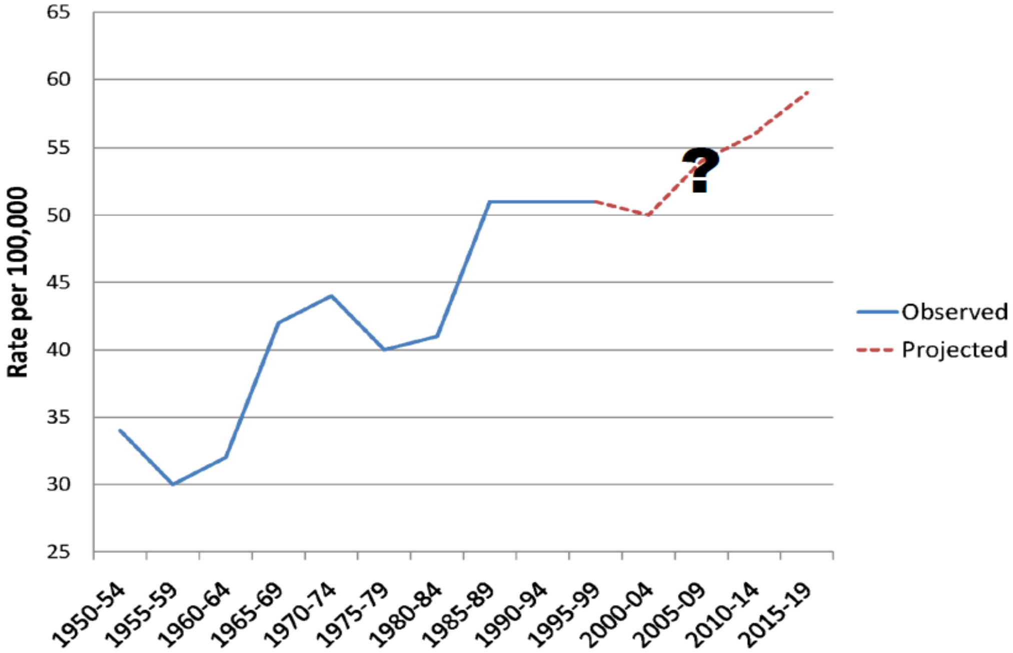
By:
Smita Kachroo , William W. Melek
University of Waterloo, Alpha Global IT
Toronto, Canada

C.J Kurian
Alpha Global IT
Toronto, Canada

Background

Accurate projections are essential for planning…

…but how best to predict future events?



Background contd..

Two basic indicators for cancer projection:

- Incidence
- Mortality

There is no uniformity in the choice of methods used for projections . With a wide range of possible methods/algorithms available, best algorithm selection is very essential to project future trends for cancer as well as to build accurate projection model.

Goal

- To design predictive models for projection of future cancer occurrences, with detailed information regarding incidence, mortality.
- To select best appropriate Data Mining/Machine Learning algorithm for developing accurate projection model.

Tools :

--- WEKA

--- Three very important algorithms(Naïve Bayes, Bayes Net and K- NN)

Why comparative analysis is important?

Many developed countries have recorded cancer incidence and mortality projections, including

1. Short term projection: usually less than 5 years ahead
2. Long term projection: around 25 years ahead

Age-Period Trend Methods for Short-Term Projections:

The global method or Average Method consider for short term projection is based on statistical regression models proposed by Bray & Møller, 2006.

Limitations:

The trends, obtained from statistical regression models, may not be reliable due to recent changes in coding, interventions (e.g. screening) .Moreover ,in his proposed method only changes of age structure and population size are taken into account to project the number of cases.

Age-Period-Cohort Approaches for Long Term projections:

Age-period-cohort models can be fitted using different approaches, such as generalized linear model(GLM) (McCullagh & Nelder, 1989), generalized additive model (GAM) (Hastie & Tibshirani,1990) and Bayesian Model.

Why comparative analysis is important? Contd…

Limitations:

- ***Bayesian age-period-cohort models*** are used increasingly to project cancer incidence and mortality rates. Data from younger age groups (typically age < 30 years) for which rates are low are often excluded from the analysis. However, empirical comparison (Baker and Bray,2005) based on data from Hungary suggests that age-specific predictions based on full data are more accurate.

- Another model proposed by Moller et al. for long-term projection of cancer incidence is from the Nordic countries called ***Nordpred model***. Nordpred model is widely used by Canadian Cancer society for projecting Canadian Cancer Statistics. The projected rates are based on the assumption that past trends will continue into the future. Any changes in these trends will mean that the projections will not be realized. This is always the case when attempting to predict future events involving uncertainty.

Why Not linear approach?

With a wide set of possible models available, as in the family of generalized linear and Non linear models, model selection is very important.

Why not Linear Models:

Limitations(to list few)

(a) In common with standard linear regression techniques, the imposition of a static model implies fixed relationships and effects across observations;

(b) The complexity of the likelihoods is such that exact inferences about observed relationships (estimation) and further observations (prediction) are precluded;

Why Non linear approach ???

In classification all decisions are based on data, but the best decisions are also based on previous experience or knowledge.

Classifiers provide a method

- For making use of previous experience in order to arrive at the best decision in interpreting data.
- Classifiers can be applied to any situation where there is uncertainty regarding the value of a measured quantity.

Data Sources

- Data used for our comparative studies is collected from Statistics Canada for the years 1981 to 2006 and is grouped by sex- male and female, twenty one cancer types and incidence / mortality counts per year. Cancer types are categorized according to the groups of ICD-O-3 and ICD-10 [5]. Remaining types are categorized as ‘others' and analyzed as a distinct set when adding up the statistics for all cancers jointly.

- We have collected the data for population from Census Canada, which give us the estimates of the total population for provinces and territories based on censuses conducted every five years from 1981 through 2006.

Validation & Comparison of a predictive learners

- There are two main facets in projection of cancer incidence and mortality: accuracy (how near to the actual value is the algorithm’s project), and efficiency (how fast can the algorithm complete the projection task)

- To compare the accuracy of our classifiers we calculated *Mean Absolute Errors, Root Mean Squared Errors, Relative Absolute Error, Precision, ROC area, TP and FP Rate for each learners.*

Validation & Comparison of a predictive model contd…

- To validate or evaluate our predictive model we used 10-fold cross-validation method.

You can use cross-validation to statistically validate the reliability of your mining model. One round of cross-validation involves partitioning a sample of data into subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

ROC Curve , Kappa and Errors measures

Kappa Statistics measures relative improvement over random predictor and is measured as

$$(\mathrm{D}_{\mathrm{observed}} - \mathrm{D}_{\mathrm{random}})/(\mathrm{D}_{\mathrm{perfect}} - \mathrm{D}_{\mathrm{random}})$$

Mean absolute error can be defined as sum of absolute errors divided by number of projections. Mean absolute error is the most popular error and is measured on actual target values: a1, a2, a3, a4….an and predicted values: p1, p2,p3, p4…pn

$$\frac{(p_1 - a_1)^2 + \cdots (p_n - a_n)^2}{n}$$

Root mean square error is defined as square root of sum of squares error divided number of predictions. It measures the differences between values predicted by a model and the values actually observed. Small value of Root mean square errors means better accuracy of model. Root mean squared error is measured as given

$$\sqrt{\frac{(\mathrm{p}_1 - \mathrm{a}_1)^2 + \cdots (\mathrm{p}_n - \mathrm{a}_n)^2}{\mathrm{n}}}$$

Relative squared error is measured as $$\frac{(p_1 - a_1)^2 + \cdots (p_n - a_n)^2}{(\bar{a} - a_1)^2 + \cdots (\bar{a} - a_n)^2}$$

Relative absolute error is measured as $$\frac{|p_1 - a_1| + \cdots |p_n - a_n|}{|(\bar{a} - a_1) + \cdots (\bar{a} - a_n)|}$$

Major Findings

The performance of Naïve Bayes, Bayesian Network and K-Nearest Neighbor was evaluated using stratified 10-fold cross-validation testing which included

- 1013 instances for incidence and 974 instances for mortality

- 4 features-Year(1981-2006), Cancer Types(Oral, Esophagus, Stomach, Colorectal, Pancreas, Larynx, Melanoma, Breast, Brain , Thyroid, Hodgkin's Lymphoma, Liver, Lung, Kidney, Multiple Myeloma, Leukemia, All , Others, Cervix, Body of Uterus, Ovary),Sex (Male and Female), and Incidence/Mortality Rates.

Major Findings

As a part of result analysis, we are isolating the correctly and incorrectly classified instances in numeric percentages. Continually we are analyzing and calculating the Kappa statistic ,ROC Area and different error rates.

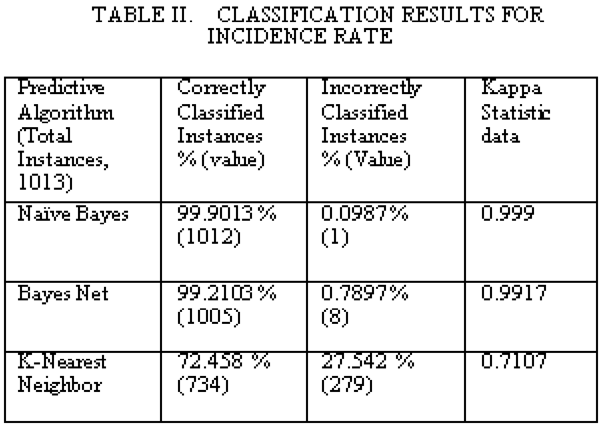


TABLE II. CLASSIFICATION RESULTS FOR INCIDENCE RATE

Predictive Algorithm (Total Instances, 1013)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Kappa Statistic data
Naïve Bayes	99.9013 % (1012)	0.0987% (1)	0.999
Bayes Net	99.2103 % (1005)	0.7897% (8)	0.9917
K-Nearest Neighbor	72.458 % (734)	27.542 % (279)	0.7107

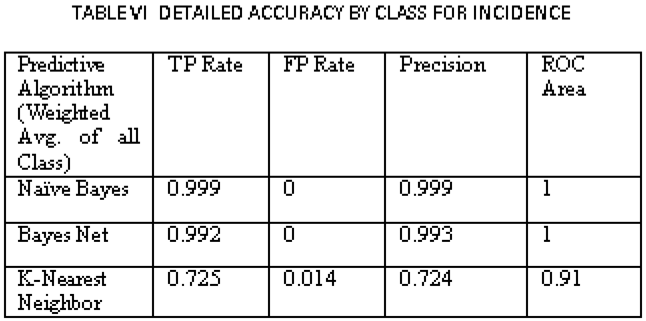


TABLE VI DETAILED ACCURACY BY CLASS FOR INCIDENCE

Predictive Algorithm (Weighted Avg. of all Class)	TP Rate	FP Rate	Precision	ROC Area
Naïve Bayes	0.999	0	0.999	1
Bayes Net	0.992	0	0.993	1
K-Nearest Neighbor	0.725	0.014	0.724	0.91

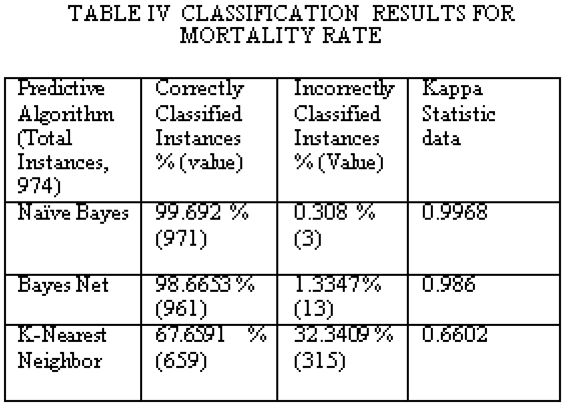


TABLE IV CLASSIFICATION RESULTS FOR MORTALITY RATE

Predictive Algorithm (Total Instances, 974)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Kappa Statistic data
Naïve Bayes	99.692 % (971)	0.308 % (3)	0.9968
Bayes Net	98.6653 % (961)	1.3347% (13)	0.986
K-Nearest Neighbor	67.6591 % (659)	32.3409 % (315)	0.6602

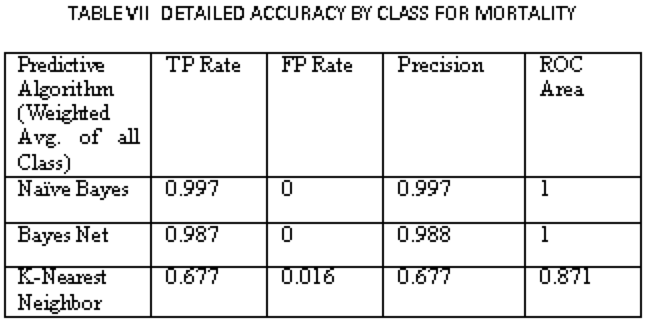


TABLE VII DETAILED ACCURACY BY CLASS FOR MORTALITY

Predictive Algorithm (Weighted Avg. of all Class)	TP Rate	FP Rate	Precision	ROC Area
Naïve Bayes	0.997	0	0.997	1
Bayes Net	0.987	0	0.988	1
K-Nearest Neighbor	0.677	0.016	0.677	0.871

Future Work

As shown in this study, with the help of certain features, predictive models can be developed that will not only supports in correctly analyzing the trends but will also benefit in accurately projecting the result of incidence and mortality rates for cancer.

The aggregated results in our paper indicated that the Naïve Bayes method achieved the greatest for incidence and mortality with classification accuracy of 99.9% which is better than any conveyed in the printed literature, the Bayes Net move towards the second best with a classification accurateness of 99.2%, and the K-Nearest Neighbor model is the last with a accuracy of 72.4%.

In our future work, we will advance our work in developing nonlinear projection model for projection of future cancer occurrences with measures like age standardized incidence and mortality rates among different sex and comparisons of the rates for a specific cancer type between several Canadian geographical regions.

Benefits of Research

Health planning, which may rely on the knowledge of what will happen in the future, is an integral part of cancer control programs. Improved health care, early detection and timely treatment is an effective approach for reducing the impact of Cancer.

- Predicting the future cancer burden is one of the first steps in knowing how to allocate resources most effectively.
- These models can help stimulate new research as well as assist decision-making and priority-setting at the individual, community, provincial/territorial and national levels.
- Provide the scope to researchers to do comparisons of cancer trends in various geographic locations.

Thank You

References

[1] Akaike H. A new look at the statistical model identification. IEEE Trans. Autom. Control. 1974;19:716–723.

[2] Bashir SA, Estève J. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. J Epidemiol Biostat 2001; 6: 287-96.

[3]Baumgartner, D.; Serpen, G. Fast Preliminary Evaluation of New Machine Learning Algorithms for Feasibility . Machine Learning and Computing (ICMLC), 2010; 113 – 115.

[4] George H. John and Pat Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.

[5] ICD10:http://www.who.int/classifications/icd/adaptations/oncology/en/

[6] Kappa at www.dmi.comumbia.edu/kappa

[7] Møller H, Fairley L, Coupland V, Okello C, Green M, Forman D, Moller B, Bray F. The future burden of cancer in England: incidence and numbers of new patients in 2020. Br J Cancer. 2007;96 (9:1484–1488).

[8] Møller B, Fekjaer H, Hakulinen T, Tryggvadottir L, Storm HH, Talback M, Haldorsen T. Prediction of cancer incidence in the Nordic countries up to the year 2020. Eur J Cancer Prev Suppl. 2002;1:S1–S96.

[9] Nils J. Nilsson (1999) Introduction to Machine Learning. California. United Stated of Americas.

[10] Osmond C. Using age, period and cohort models to estimate future mortality rates. Int J Epidemiol. 1985; 14(1):124-9.

[11] Stewart P & Xie L. Results of survey on cancer projection methods used in the provincial /territorial cancer registries. Presentation at Canadian Cancer Statistics Steering Committee Meeting, Toronto, June 23, 2009

[12] T. Darrell and P. Indyk and G. Shakhnarovich (2006). Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press.

[13] Weka 3: Data mining with open source machine learning software in java. http://www.cs.waikato.ac.nz/ ml/weka/.

[14] Wolpert DH, Macready WG, David H, William G. No Free Lunch Theorems for Search. Technical Report SFI-TR-95-02-010 (Santa Fe Institute) 1995

[15] Mathers CD, Loncar D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. PLoS Medicine 2006;3(11):e442.

[16] Warren J. Cancer death rates falling, but slowly. WebMD medical news;2003(http://aolsvc.health.webmd.aol.com/content/Artcile/73/82013.htm)

[17] Progress shown in death rates from four leading cancers (http://cancer.gov/newscenter/pressreleases/2003 Report Release).

[18] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Wermter S, Riloff E, Scheler G, editors. The Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)San Francisco, CA: Morgan Kaufman; 1995. p. 1137—45.

[19] Baker A, Bray I. Bayesian projections: what are the effects of excluding data from younger age groups?Am J Epidemiol 2005; 162: 798-805.

[20] Bray F and Møller B. Predicting the future burden of cancer. Nature Reviews, 2006, 6: 63 - 74.

[21] Burke HB, Goodman PH, Rosen DB, Henson DE, WeinsteinJN, Harrell Jr FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:857—62.

[22] Fawcett, T. “An introduction to ROC analysis”; Pattern Recognition Letters, Vol. 27 Issue 8, pp. 861-874, 2006.

[23] Data Mining: Practical Machine Learning Tools and Techniques (Chapter 5), Ian H. Witten, Eibe Frank, Mark A. Hall.

[24] Howlader N, Noone AM, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2010, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2010/, based on November 2012 SEER data submission, posted to the SEER web site, April 2013.