Journal of Physics: Conference Series **490** (2014) 012135 doi:10.1088/1742-6596/490/1/012135

Computational diagnosis of canine lymphoma

E M Mirkes[1], I Alexandrakis[2], K Slater[3], R Tuli[2] and A N Gorban[1]

[1] Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK
[2] Avacta Animal Health, Unit 706, Avenue E, Thorp Arch Estate Wetherby, LS23 7GA
[3] PetScreen Ltd, Biocity, Pennyfoot Street, Nottingham, NG1 1GF, UK

E-mail: ag153@le.ac.uk

Abstract. One out of four dogs will develop cancer in their lifetime and 20% of those will be lymphoma cases. PetScreen developed a lymphoma blood test using serum samples collected from several veterinary practices. The samples were fractionated and analysed by mass spectrometry. Two protein peaks, with the highest diagnostic power, were selected and further identified as acute phase proteins, C-Reactive Protein and Haptoglobin. Data mining methods were then applied to the collected data for the development of an online computer-assisted veterinary diagnostic tool. The generated software can be used as a diagnostic, monitoring and screening tool. Initially, the diagnosis of lymphoma was formulated as a classification problem and then later refined as a lymphoma risk estimation. Three methods, decision trees, kNN and probability density evaluation, were used for classification and risk estimation and several pre-processing approaches were implemented to create the diagnostic system. For the differential diagnosis the best solution gave a sensitivity and specificity of 83.5% and 77%, respectively (using three input features, CRP, Haptoglobin and standard clinical symptom). For the screening task, the decision tree method provided the best result, with sensitivity and specificity of 81.4% and >99%, respectively (using the same input features). Furthermore, the development and application of new techniques for the generation of risk maps allowed their user-friendly visualization.

1. Introduction

Lymphoma (Lymphosarcoma, LSA) is one of the most common cancers seen in dogs. One in four dogs will develop cancer in their lifetime. It accounts for approximately 20% of all canine tumours [1].

The PetScreen Canine Lymphoma Blood Test employs advanced technology to detect lymphoma biomarkers present in a dog's serum [1]. Concentration of two acute phase proteins is evaluated: Haptoglobin (Hapt) and C-Reactive Protein (CRP) Detection of these biomarkers indicates a high likelihood that the dog has lymphoma [1], [2]. The growth of the database enables new methods of data mining. We analyse usability of attributes for the lymphoma diagnostic test. We divide the database in two cohorts and formulate two different tasks: (i) *differential diagnostic in clinically suspected cases* and (ii) *screening*. The isolation of the clinically suspected cohort is necessary for formulation of the task of differential diagnostics and selection of the appropriate methods.

Three methods are used. The first is the method of *decision trees* [3-5]. The second is *k nearest neighbours method* (kNN) in several versions [6,7]. The third is the method of *probability density function estimation* (PDFE) [8,9]. We use them for direct estimation of the lymphoma risk.

We present the case study for both tasks: for the diagnostics task we have tested 2,432,000 variants of the kNN method, 248,400 variants of decision tree algorithms and 280 variants of PDFE method; for the screening task we have tested 48,640 variants of kNN, 4,968 variants of decision trees and 280 variants of PDFE.

The versions differ by impurity criteria, kernel functions, number of nearest neighbours, weights and other parameters. They are compared by the standard goodness of fit, the entropic criteria and the generalization ability.

The best results are implemented in web-accessed software for the diagnosis of canine lymphoma.

If we solve the classification problem then for each value of input features we can obtain only two answers: dog with lymphoma or dog without lymphoma. Formulation of the task as a problem of risk estimation provides obtaining the real value between 0 and 1. Analysis of risk map provides a new ability to formulate hypotheses.

The obtained results provide the creation of a reliable diagnostic and screening system for canine lymphoma.

2. Data description

Table 1. Relative information gain for 'Lymphoma'

Tested feature	RIG	RIG under given Lymphadenopathy	
		Y	N
Lymphadenopathy	28.92%	-	-
CRP binned	24.38%	15.00%	23.52%
Hapt binned	7.02%	1.76%	14.32%
Age binned	6.07%	1.62%	9.39%
Sex	0.95%	3.79%	22.84%

Categorical features. Lymphoma ('Y' or 'N'), Sex ('M' or 'F'), and a most important clinical symptom, Lymphadenopathy ('Y' or 'N')

Real features. Age, CRP (concentration), Hapt (concentration).

Analysis of importance. Table 1 contains values of relative information gain (RIG) [3] for target feature Lymphoma of any input feature. RIG is calculated for whole database and for two samples: (Y) with Lymphadenopathy='Y' and (N) with Lymphadenopathy='N'.

3. Problem refinement

The results of the database analysis show that there are two different cohorts of data in the database.

The first cohort is titled 'clinically suspected' and contains records collected by PetScreen from dogs undergoing differential diagnosis. All these records correspond to dogs which have been referred to differential diagnostics by veterinary practitioners on the base of one or more clinical symptoms. It is not possible to find a posteriori these symptoms for each instance and we have to introduce a new synthetic attribute: 'clinically suspected'. The second cohort is titled 'healthy' and contains records obtained from healthy dogs courtesy of the Pet Blood Bank. These two cohorts have very different statistics of the attributes. By expert estimations, the prior probability of lymphoma is located between 2% and 5% in the canine population. The number of records of dogs with lymphoma is 97 or 32% of all the records in the initial database. All these cases have been clinically suspected and form 42% of the clinically suspected cases. This imbalance entails the use of specific methods to solve screening tasks. The 'clinically suspected' feature was added to the database to identify the two cohorts. The values of feature 'clinically suspected' were defined by using additional information of veterinary cards. The existence of the two cohorts allows the formulating of two different tasks: differential diagnostics and screening.

Differential diagnostics. The differential diagnostic task can be formulated as a problem of lymphoma diagnosis for dogs with some clinical symptoms of lymphoma. A diagnostic task is a usual classification problem and all classification methods can be used. We use three types of classification methods: kNN, decision tree and the method based on probability distribution function estimation. The first two methods have an additional parameter – weighting of the positive class.

Screening. The screening task can be formulated as an estimate of lymphoma risk for any dog. To solve this task we used all the database records. The experts' estimation of prior probability of lymphoma is between 2% and 5% however the fraction of dogs with lymphoma records in the database is 32%. To compensate for this imbalance all methods take into account the prior probability of lymphoma. For this task, the weighting of classes is defined by prior probability.

We use the following notations: p is the prior probability of lymphoma, N_L is the number of dogss with lymphoma, N_{CS} is the number of all clinically suspected dogs and N_H is the number of healthy dog. The weight of the class of dogs with lymphoma is equal to p. The weight of one dog with lymphoma is equal to $w_L = p/N_L$. Indeed, this is the weight of any record of the clinically suspected cohort. The weight of each record of a healthy dog is calculated as $w_H = (1\text{-}w_L N_{CS})/N_H$. These two weights are used to calculate the risk of lymphoma in the screening task.

4. Risk map analysis

Visualization of data and probability distributions may use various screens, from the coordinate planes and PCA to non-linear principal graphs and manifolds [10]. In this work, we use visualisation of risk of lymphoma on the plane of two real attributes, CRP and Hapt. For example, we use risk maps to generate hypotheses about impact of input features. All maps below use the legend which is presented in any figure. The brown colour indicates the low level of risk: the greater intensity of brown colour indicates the less risk of lymphoma. The white colour indicates the median value of risk (50%). Blue colour indicates the high level of risk: the greater intensity of indicates the greater risk value.

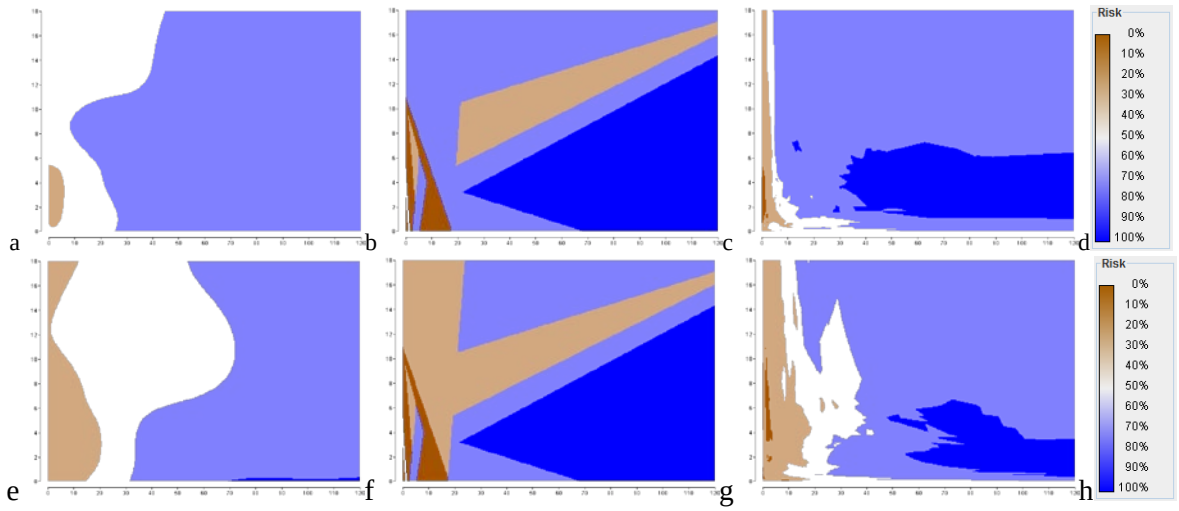

Figure 1. The maps of lymphoma risk for male and female dogs: a) PDFE map for male, b) decision tree map for male, c) KNN map for male, e)PDFE map for female, f) decision tree map for female, g) KNN map for female, d) and h) are legend.

Let us consider the risk of lymphoma in relation to sex for clinically suspected dogs. The probabilities are represented (Fig 1). The maps show that there are some areas where the risk of lymphoma is greater for male dogs. For the best decision tree and kNN the qualitatively same results were obtained (Fig. 2). In this area, the risk of lymphoma may depend on the steroid hormones. This hypothesis needs additional verification.

5. Results

During the case-study the best solution for each task is selected with the minimal number of errors.

Differential diagnostic. The best result is obtained by a decision tree which uses three input features: the concentrations of CRP and Hapt, and Lymphadenopathy. The concentrations of CRP and

Hapt are used in linear combinations. DKM is used as a splitting criterion. The sensitivity of this method is 83.5%, specificity is 77%. ROC integral for this method is 0.879.

If 'Lymphadenopathy' is unknown then we use decision tree which only uses CRP and Hapt. In the best version, these input features are used in linear combinations after logarithmic transformation. Information gain is used as the splitting criterion. The sensitivity of this method is 81.5%, specificity is 76%. ROC integral for this method is 0.810.

Screening. The best result is obtained by the decision tree which uses three input features: the concentrations of CRP and Hapt, and Lymphadenopathy. CRP and Hapt are used separately. DKM is used as the splitting criterion. The sensitivity of this method is 81.4%, specificity is >99% (no false negative results in one-leave-out cross-validation).

If 'Lymphadenopathy' is unknown then we use another decision tree, which uses CRP and Hapt only. These input features are used in linear combinations after logarithmic transformation. Gini gain is used as the splitting criteria. The sensitivity is 65%, specificity is 83%.

To evaluate the quality of the achieved results, we compare them to some current human cancer screening tests. The tests that rely upon single biomarkers demonstrate often the worse performance. For example, the CA-125 screen for human ovarian cancer provides sensitivity approximately 53% and specificity 98%, and the male PSA test gives sensitivity approximately 85% and specificity 35%. Supplementation of CA-125 by several other biomarkers increases sensitivity of at least 75% for early stage disease and specificity of 99.7%. For PSA marker, using age-specific reference ranges improved test specificity and sensitivity, but did not improve the overall accuracy of PSA testing.

Visualisation of risk maps provides a friendly tool for explanatory data analysis and affords an opportunity to generate hypotheses about impact of input feature on the final diagnosis. For more details and the additional bibliography we refer to [11].

References

[1] Ratcliffe L, Mian S, Slater K, King H, Napolitano M, Aucoin D and Mobasheri A 2009 Proteomic identification and profiling of canine lymphoma patients, *Veterinary and Comparative Oncology* **7(2)** pp 92–105

[2] Alexandrakis I 2012 The use of CART algorithms to combine serum acute phase protein levels as a diagnostic aid in canine lymphoma *Proc. of 15th Congress of the Int. Society for Animal Clinical Pathology, 14th Conf. of the European Society of Veterinary Clinical Pathology, (Ljubljana, Slovenia, 3rd-7th July, 2012)* (Ljubljana: Veterinary Faculty) p 65

[3] Rokach L and Maimon O 2010 Decision trees *Data Mining and Knowledge Discovery Handbook* ed O Maimon and L Rokach (Berlin: Springer) pp 165–192

[4] Quinlan J R 1987 Simplifying decision trees *Int. J. of Man-Machine Studies* **27** pp 221–234

[5] Gelfand S B, Ravishankar C S and Delp E J 1991 An iterative growing and pruning algorithm for classification tree design *IEEE Transaction on Pattern Analysis and Machine Intelligence* **13(2)** pp 163–174

[6] Clarkson K L 2005 Nearest-neighbor searching and metric space dimensions *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (The MIT Press) pp 15–59.

[7] Hastie T and Tibshirani T 1996 Discriminant adaptive nearest neighbor classification *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18 (6)** pp 607–616

[8] Scott D W 1992 *Multivariate Density Estimation: Theory, Practice and Visualization* (New York: Wiley)

[9] Buhmann M D 2003 *Radial Basis Functions: Theory and Implementations* (Cambridge University Press)

[10] Gorban A N, Kégl B, Wunsch D C and Zinovyev A (eds) 2007 *Principal Manifolds for Data Visualisation and Dimension Reduction* (*LNCSE*) vol 58 (Berlin – Heidelberg – New York: Springer)

[11] Mirkes E M, Alexandrakis I, Slater K, Tuli R and Gorban A N 2023 arXiv:1305.4942 [q-bio.QM]