UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/11/06

The Fisher Kernel: A Brief Review

20 January 2011

Martin Sewell

Abstract

The basic idea behind the Fisher kernel method is to train a (generative) hidden Markov model (HMM) on data to derive a Fisher kernel for a (discriminative) support vector machine (SVM). The Fisher kernel gives a 'natural' similarity measure that takes into account the underlying probability distribution. If each data item is a (possibly varying length) sequence, each may be used to train a HMM, with the average of the models in the training set used to construct a global HMM. It is then possible to calculate how much a new data item would 'stretch' the parameters of the existing model. This is achieved by, for two data items, calculating and comparing the gradient of the log-likelihood of the data item with respect to the model with a given set of parameters. If these 'Fisher scores' are similar it means that the two data items would adapt the model in the same way, that is from the point of view of the given parametric model at the current parameter setting they are similar in the sense that they would require similar adaptations to the parameters. This brief note introduces the Fisher score and the Fisher kernel, then reviews the literature on Fisher kernels.

1 Definitions

Let $P(x|\boldsymbol{\theta})$ be a probability model, where x is a data item and $\boldsymbol{\theta}$ is a vector of the model parameters. $\nabla_{\boldsymbol{\theta}}$ is the gradient operator with respect to $\boldsymbol{\theta}$, and $\log_e P(x|\boldsymbol{\theta})$ is the log-likelihood of x with respect to the model with a given set of parameters $\boldsymbol{\theta}$. Then the Fisher score, U_x, is the gradient of the log-likelihood of x with respect to the model with a given set of parameters $\boldsymbol{\theta}$.

$$U_x = \nabla_\theta \log_e P(x|\theta)$$

The Fisher score gives us an embedding into the feature space $\mathbb{R}^N$. The Fisher kernel refers to the inner product in this space, and is defined as

$$K(x_i, x_j) = U_{x_i}^T I^{-1} U_{x_j}$$

where I is the Fisher information matrix. The Fisher kernel engenders a measure of similarity between two data items x_i and x_j by defining a distance between them, and can be used with any kernel-based classifier such as a support vector machine (SVM).

2 Literature Review

Jaakkola and Haussler (1999a) introduced the Fisher kernel (named in honour of Sir Ronald Fisher), thus creating a generic mechanism for incorporating generative probability models into discriminative classifiers such as SVMs. Jaakkola and Haussler (1999b) introduced a generic class of probabilistic regression models and a parameter estimation technique that can make use of arbitrary kernel functions. Jaakkola et al. (1999) applied the Fisher kernel method to detecting remote protein homologies which performed well in classifying protein domains by SCOP (Structural Classification of Proteins) superfamily. Jaakkola et al. (2000) found that using the Fisher kernel significantly improved on previous methods for the classification of protein domains based on remote homologies. Moreno and Rifkin (2000) used the Fisher kernel method for large scale Web audio classification. Smith and Niranjan (2000) give some experimental justification for the Fisher kernel. They explain that Fisher kernels limit the dimension of the feature space, and their results suggest that limiting the feature space dimension may give some beneficial regularization, particularly when the two classes are very inseparable. They also discuss the fact that when data is costly to label, Fisher kernels provide a means of using both the labelled and unlabelled data. Fine et al. (2001) built a hybrid Gaussian mixture model (GMM)/Fisher kernel SVM and applied it to speaker identification, with positive results. Vinokourov and Girolami (2001) successfully employed the Fisher kernel for document classification. Wan and Renals (2002) used SVMs for speaker verification and identification tasks. They compared the polynomial kernel, the Fisher kernel, a likelihood ratio kernel and the pair hidden Markov model (HMM) kernel with baseline systems trained on a discriminative polynomial classifier and generative GMM classifiers. The Fisher kernel performed well. Wan and Renals (2003) applied SVMs to speaker verification using the Fisher kernel and the likelihood ratio kernel. Saunders et al. (2003) showed how the string kernel can be thought of as a k-stage Markov process, and as a result interpreted as a Fisher kernel. Tsuda et al. (2004) analysed the statistical properties of the Fisher kernel. Nicotra et al. (2004) extended the Fisher kernel to deal with tree structured data. Kersting and Gärtner (2004) extended the Fisher kernel to logical sequences (sequences over an alphabet of logical atoms). Their experiments showed a considerable improvement over results achieved without Fisher kernels for logical sequences.

Wan and Renals (2005) present a text-independent speaker verification system using SVMs with 'score-space kernels'—kernels that generalize Fisher kernels and are based on underlying generative models such as GMMs. Holub et al. (2005) successfully combined generative models with Fisher kernels to realize performance gains on standard object recognition data sets. Elkan (2005) investigated the Dirichlet compound multinomial (DCM) Fisher kernel (the Fisher kernel induced by the DCM distribution). His experiments showed that the DCM Fisher kernel performed better than alternative kernels for nearest-neighbour document classification, but that the term frequency-inverse document frequency (TF-IDF) representation still performed best. Dick and Kersting (2006) developed Fisher kernels for relational data. Gunsel et al. (2006) compared the classification performance of HMMs, SVMs and an SVM with a Fisher kernel on a small database of two-handed gestures. On average, the HMM performed best, but the single best result was due to the Fisher kernel. Perronnin and Dance (2007) applied the Fisher kernel framework to a visual vocabulary, i.e. a GMM which models the generative process of the low-level feature vectors extracted from images. They showed that the proposed approach is actually a generalization of the popular bag-of-visterms. In a new method for topic-based text segmentation, Sun et al. (2008) introduced a latent Dirichlet allocation (LDA)-based Fisher kernel to exploit text semantic similarities, then employed dynamic programming to obtain global optimization. Chappelier and Eckard (2009) (1) introduced a new, rigorous development of the Fisher kernel for the Probabilistic Latent Semantic Indexing model (PLSI), addressing the role of the Fisher Information Matrix, and uncovering its relation to the kernels proposed so far; and (2) proposed a novel and theoretically sound document similarity measure, which avoids the problem of 'folding in' unknown documents. Their overall conclusion, however, was that

PLSI is not well-suited for large scale ad hoc information retrieval (IR), mainly because it is not a fully generative model. Aran and Akarun (2010) proposed a multi-class classification strategy that applies a multi-class classification on each Fisher score space and combines the decisions of multi-class classifiers. They showed experimentally that the Fisher scores of one class provide discriminative information for the other classes as well. The authors applied their technique to a sign language data set with significant success.

References

Aran, O. and Akarun, L. (2010), A multi-class classification strategy for Fisher scores: Application to signer independent sign language recognition, *Pattern Recognition* **43**(5), 1776–1788.

Chappelier, J.-C. and Eckard, E. (2009), PLSI: The true Fisher kernel and beyond: IID Processes, information matrix and model identification in PLSI, *in* W. Buntine, M. Grobelnik, D. Mladenić and J. Shawe-Taylor (eds.), *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 2009, Proceedings, Part I*, Vol. 5781 of *Lecture Notes in Computer Science*, Springer, Berlin, pp.195–210.

Dick, U. and Kersting, K. (2006), Fisher kernels for relational data, *in* J. Fürnkranz, T. Scheffer and M. Spiliopoulou (eds.), *Machine Learning: ECML 2006: 17th European Conference on Machine Learning, Berlin, Germany, September 2006, Proceedings*, Vol. 4212 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, pp.114–125.

Elkan, C. (2005), Deriving TF-IDF as a Fisher kernel, *in* M. Consens and G. Navarro (eds.), *String Processing and Information Retrieval: 12th International Conference, SPIRE 2005 Buenos Aires, Argentina, November 2005. Proceedings*, Vol. 3772 of *Lecture Notes in Computer Science*, Springer, Berlin, pp.295–300.

Fine, S., Navrátil, J. and Gopinath, R. A. (2001), A hybrid GMM/SVM approach to speaker identification, *in The 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. Volume I*, IEEE, Piscataway, NJ, pp.417–420.

Gunsel, B., Jain, A. K., Tekalp, A. M. and Sankur, B. (2006), Recognizing two handed gestures with generative, discriminative and ensemble methods via Fisher kernels, *in* B. Gunsel, A. K. Jain, A. M. Tekalp and B. Sankur (eds.), *Multimedia Content Representation, Classification and Security: International Workshop, MRCS 2006, Istanbul, Turkey, September 2006. Proceedings*, Vol. 4105 of *Lecture Notes in Computer Science*, Springer, Berlin, pp.159–166.

Holub, A. D., Welling, M. and Perona, P. (2005), Combining generative models and Fisher kernels for object recognition, *in Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Volume 1*, IEEE, Washington, DC, pp.136–143.

Jaakkola, T., Diekhans, M. and Haussler, D. (1999), Using the Fisher kernel method to detect remote protein homologies, *in* T. Lengauer, R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes and R. Zimmer (eds.), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp.149–158.

Jaakkola, T., Diekhans, M. and Haussler, D. (2000), A discriminative framework for detecting remote protein homologies, *Journal of Computational Biology* **7**(1–2), 95–114.

Jaakkola, T. S. and Haussler, D. (1999a), Exploiting generative models in discriminative classifiers, *in* M. S. Kearns, S. A. Solla and D. A. Cohn (eds.), *Advances in Neural Information Processing Systems 11*, Bradford Books, The MIT Press, Cambridge, MA, pp.487–493.

Jaakkola, T. S. and Haussler, D. (1999b), Probabilistic kernel regression models, *in* D. Heckerman and J. Whittaker (eds.), *Proceedings of the 1999 Conference on AI and Statistics*, Morgan Kaufmann, San Mateo, CA.

Kersting, K. and Gärtner, T. (2004), Fisher kernels for logical sequences, *in* J.-F. Boulicaut, F. Esposito, F. Giannotti and D. Pedreschi (eds.), *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 2004. Proceedings*, Vol. 3201 of *Lecture Notes in Computer Science*, Springer, Berlin, pp.205–216.

Moreno, P. J. and Rifkin, R. (2000), Using the Fisher kernel method for Web audio classification, *in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings, Volume IV*, IEEE, pp.2417–2420.

Nicotra, L., Micheli, A. and Starita, A. (2004), Fisher kernel for tree structured data, *in* J. C. Principe, D. C. Wunsch, M. Hasselmo, F. M. Ham, D. G. Brown, G. Carpenter, D. Casasent, D. Floreano, K. Fukushima, S. Gielen, S. Grossberg, B. Kosko, R. Kozma, G. Lendaris, D. S. Levine, W. Levy, F. C. Morabito, K. Obermayer, E. Oja, D. Prokhorov, H. Szu, J. G. Taylor, P. Werbos, B. Widrow, D. Wang and L. A. Zadeh (eds.), *Proceedings. 2004 IEEE International Joint Conference on Neural Networks. Volume 3*, IEEE, pp.1917–1922.

Perronnin, F. and Dance, C. (2007), Fisher kernels on visual vocabularies for image categorization, *in IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, pp.1–8.

Saunders, C., Shawe-Taylor, J. and Vinokourov, A. (2003), String kernels, fisher kernels and finite state automata, *in* S. Becker, S. Thrun and K. Obermayer (eds.), *Advances in Neural Information Processing Systems 15*, Bradford Books, The MIT Press, Cambridge, MA, pp.633–640.

Smith, N. and Niranjan, M. (2000), Data-dependent kernels in SVM classification of speech patterns, *in Proceedings, Sixth International Conference on Spoken Language Processing (ICSLP 2000), vol. 1*, pp.297–300.

Sun, Q., Li, R., Luo, D. and XihongWu (2008), Text segmentation with LDA-based Fisher kernel, *in 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Proceedings of the Conference*, Association for Computational Linguistics, Stroudsburg, pp.269–272.

Tsuda, K., Akaho, S., Kawanabe, M. and Müller, K.-R. (2004), Asymptotic properties of the Fisher kernel, *Neural Computation* **16**(1), 115–137.

Vinokourov, A. and Girolami, M. (2001), Document classification employing the Fisher kernel derived from probabilistic hierarchic corpus representations, *in Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, British Computer Society Information Retrieval Specialist Group, Springer-Verlag, Berlin, pp.24–40.

Wan, V. and Renals, S. (2002), Evaluation of kernel methods for speaker verification and identification, *in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. Vol. I*, IEEE, pp.669–672.

Wan, V. and Renals, S. (2003), SVMSVM: Support vector machine speaker verification methodology, *in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. Vol. II*, IEEE, Piscataway, NJ, pp.221–224.

Wan, V. and Renals, S. (2005), Speaker verification using sequence discriminant support vector machines, *IEEE Transactions on Speech and Audio Processing* **13**(2), 203–210.