10.12871/CLICIT2014169

Are Quantum Classifiers Promising?

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. This paper presents work in progress on the development of a new general purpose classifier based on Quantum Probability Theory. We will propose a kernel-based formulation of this classifier that is able to compete with a state-of-the-art machine learning methods when classifying instances from two hard artificial problems and two real tasks taken from the speech processing domain.

Italiano. *Questo contributo presenta i primi risultati di un progetto per lo sviluppo di un classificatore basato sulla teoria della probabilità quantistica. Presenteremo un modello basato su kernel in grado di competere con i migliori metodi di machine learning considerando i due problemi artificiali complessi e i due casi reali sui quali è stato valutato.*

1 Introduction

Quantum Mechanics Theory (QMT) is one of the most successful theory in modern science. Despite its ability to properly describe most natural phenomena in the physics realm, the attempts to prove its effectiveness in other domains remain quite limited. Only in recent years some scholars tried to embody principles derived from QMT into their specific fields. This connection has been actively studied, for example, by the Information Retrieval community (Zuccon *et al.*, 2009; Melucci, van Rijsbergen, 2011; Gonzàlez, Caicedo, 2011) and in the domain of cognitive sciences and decision making (Busemeyer, Bruza, 2012). Also the NLP community started to look at QMT with interest and some studies using it have already been presented (Blacoe *et al.*, 2013; Liu *et al.*, 2013).

This paper presents work in progress on the development of a new classifier based on Quantum Probability Theory. Starting from the work presented in (Liu *et al.*, 2013) we will show all the limits of this simple quantum classifier and propose a new kernel-based formulation able to solve most of its problems and able to compete with a state-of-the-art classifier, namely Support Vector Machines, when classifying instances from two hard artificial problems and two real tasks taken from speech processing domain.

2 Quantum Probability Theory

A *quantum state* denotes an unobservable distribution which gives rise to various observable physical quantities (Yeang, 2010). Mathematically it is a vector in a complex Hilbert space. It can be written in Dirac notation as $|\psi\rangle = \sum_1^n \lambda_j \, |e_j\rangle$where λ_j are complex numbers and the $|e_j\rangle$ are the basis of the Hilbert space ($|.\rangle$ is a column vector, or a *ket*, while $\langle.|$ is a row vector, or a *bra*). Using this notation the inner product between two state vectors can be expressed as $\langle\psi|\phi\rangle$ and the outer product as $|\psi\rangle \langle\phi|$.

$|\psi\rangle$ is not directly observable but can be probed through measurements. The probability of observing the elementary event $|e_j\rangle$ is $| \langle e_j|\psi\rangle |^2 = |\lambda_j|^2$ and the probability of $|\psi\rangle$ collapsing on $|e_j\rangle$ is $P(e_j) = |\lambda_j|^2 / \sum_1^n |\lambda_i|^2$ (note that $\sum_1^n |\lambda_i|^2 = \| |\psi\rangle \|^2$ where $\|\cdot\|$ is the vector norm). General events are subspaces of the Hilbert space.

A matrix can be defined as a *unitary operator* if and only if $UU^\dagger = I = U^\dagger U$, where † indicates the Hermitian conjugate. In quantum probability theory unitary operators can be used to evolve a quantum system or to change the state/space basis: $|\psi'\rangle = U \, |\psi\rangle$.

Quantum probability theory (see (Vedral, 2007) for a complete introduction) extends standard kolmogorovian probability theory and it is in principle adaptable to any discipline.

3 Quantum Classifiers

3.1 The Classifier by (Liu *et al.*, 2013)

In their paper Liu *et al.* presented a quantum classifier based on the early work of (Chen, 2002). Given an Hilbert space of dimension $n = n_i + n_o$, where n_i is the number of input features and n_o is the number of output classes, they use a unitary operator U to project the input state contained in the subspace spanned by the first n_i basis vectors into an output state contained in the subspace spanned by the last n_o basis vectors: $|\psi^o\rangle = U|\psi^i\rangle$. Input, $|\psi^i\rangle$, and output, $|\psi^o\rangle$, states are real vectors, the former having only the first n_i components different from 0 (assigned to the problem input features of every instance) and the latter only the last n_o components. From $|\psi^o\rangle$ they compute the probability of each class as $P(c_j) = |\psi^o_{ni+j}|^2 / \sum_1^{no} |\psi^o_{ni+i}|^2$ for $j = 1..n_o$.

The unitary operator U for performing instances classification can be obtained by minimising the loss function

$$err(T) = 1/\sum_{j=1}^{|T|} \langle \psi_j^o | \psi_j^t \rangle \,,$$

where T is the training set and $|\psi^t\rangle$ is the target vector for output probabilities (all zeros except 1 for the target class) for every instance k, using standard optimisation techniques such as Conjugate Gradient (Hestenes, Stiefel, 1952), L-BFGS (Liu, Nocedal, 1989) or ASA (Ingber, 1989).

This classifier exhibits interesting properties. Let us examine its behaviour by using a standard non-linear problem: the XOR problem. The four instances of this problem are:

$$\begin{array}{ll}
|\psi_1^i\rangle = (-1,-1,0,0) & |\psi_1^t\rangle = (0,0,1,0) \\
|\psi_2^i\rangle = (-1,1,0,0) & |\psi_2^t\rangle = (0,0,0,1) \\
|\psi_3^i\rangle = (1,-1,0,0) & |\psi_3^t\rangle = (0,0,0,1) \\
|\psi_4^i\rangle = (1,1,0,0) & |\psi_4^t\rangle = (0,0,1,0)
\end{array}$$

Figure 1 depicts the probability functions for both classes as well as the decision boundaries where $P(c_1) > P(c_2)$ after a training session. Despite the relative simplicity of this classifier the two probability functions are non-linear, but the decision boundaries are linear. Nevertheless it is able to correctly classify the instances of the XOR problem.

The simplicity and the low power of this classifier emerge quite clearly when we test it with more difficult, though linearly separable, classification problems. Figure 2 shows the results of the (Liu *et al.*, 2013) classifier when applied to two simple problems. In both cases the classifier is not able to properly divide the input space into different regions corresponding to the required classes. Moreover, all the decision boundaries have to cross the origin of the feature space, a very limiting constraint for general classification problems, and problems that require strict non-linear decision boundaries cannot be successfully handled by this classifier. Nevertheless the ability of managing a classical non-linear problem, the XOR problem, is very promising and extending this method could lead, in our opinion, to interesting results.

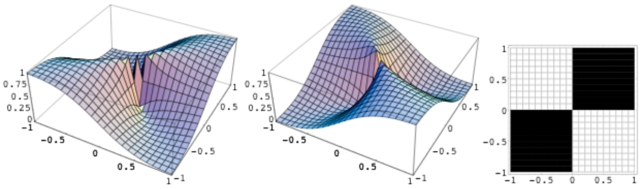

Figure 1: The probability functions for c_1 (left) and c_2 (center) for the XOR problem. At right, the decision boundaries between the two classes, where $P(c_1) > P(c_2)$ is marked in black.

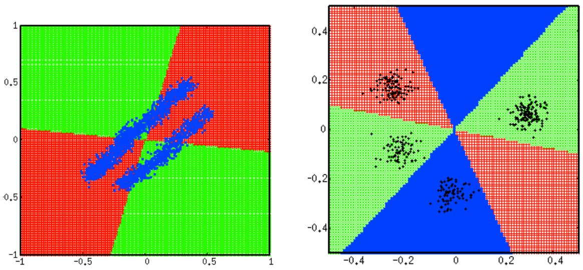

Figure 2: A two-class problem (left) and a four-class problem that cannot be successfully handled by the classifier proposed by (Liu *et al.*, 2013).

3.2 Kernel Quantum Classifier (KQC)

The goal of this paper is to extend the examined classifier in various direction in order to obtain a classification tool with higher performances.

A widely used technique to transform a linear classifier into a non-linear one involves the use of the “kernel trick”. A non-linearly separable problem in the input space can be mapped to a higher-dimensional space where the decision borders between classes might be linear. We can do that through the mapping function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, with $m > n$, that maps an input state vector $|\psi^i\rangle$ to a new space. The interesting thing is that in the new space, for some particular mappings, the inner product can be calculated by using *kernel* functions $k(x,y) = \langle\phi(x),\phi(y)\rangle$ without explicitly computing the mapping ϕ of the two original vectors.

We can express the unitary operator performing the classification process as a combination of the training input vectors in the new features space

$$\begin{aligned} |\psi^o\rangle &= U\ |\phi(\psi^i)\rangle \\ |\psi^o\rangle &= \textstyle\sum_{j=1}^{|T|} |\alpha_j\rangle\ \langle\phi(\psi_j^i)|\ \ |\phi(\psi^i)\rangle \\ |\psi^o\rangle &= \textstyle\sum_{j=1}^{|T|} |\alpha_j\rangle\ \langle\phi(\psi_j^i)|\phi(\psi^i)\rangle \end{aligned}$$

that can be rewritten using the kernel as

$$|\psi^o\rangle = \sum_{j=1}^{|T|} |\alpha_j\rangle\, k(\psi_j^i,\psi^i). \qquad (1)$$

Adding a bias term $|\alpha_0\rangle$ to the equation (1) lead to the final model governing this new classifier:

$$|\psi^o\rangle = |\alpha_0\rangle + \sum_{j=1}^{|T|} |\alpha_j\rangle\, k(\psi_j^i,\psi^i) \qquad (2)$$

In this new formulation we have to obtain all the $|\alpha_j\rangle$ vectors, $j = 0,...,|T|$, through an optimisation process similar to the one of the previous case, minimising a standard euclidean loss function

$$err(T) = \sum_{j=1}^{|T|} \sum_{k=1}^{no} \left(P_j(c_k) - \psi^t_{j(ni+k)}\right)^2 + \gamma \sum_{j=0}^{|T|} \||\alpha_j\rangle\|.$$

using a numerical optimisation algorithm, L-BFGS in our experiments, where $P(c)$ is the class probability defined in section 3.1 and $\gamma\sum\||\alpha_j\rangle\|$ is an L_2 regularisation term on model parameters (the real and imaginary parts of $|\alpha_j\rangle$ components).

Once learned a good model from the training set T, represented by the $|\alpha_j\rangle$ vectors, we can use equation (2) and the definition of class probability for classifying new instance vectors.

It is worth noting that the KQC proposed here involves a large number of variables during the optimisation process (namely, $2 * no * (|T| + 1)$) that depends linearly on the number of instances in the training set T. In order to build a classifier applicable to real problems, we have to introduce special techniques to efficiently compute the gradient needed by optimisation methods. We relied on Automatic Differentiation (Griewank, Walther, 2008), avoiding any gradient approximation using finite differences that would require a very large number of error function evaluations. Using such techniques the training times of KQC are comparable to those of other machine learning methods.

Figure 3a and 3b show the classification results of KQC, using the linear kernel ($k(x,y) = \langle x,y\rangle$), when applied to the same problems analysed before to describe the behaviour of the (Liu *et al.*, 2013) classifier. KQC is able to discriminate efficiently between linearly separable binary or multiclass problems adapting the decision boundaries in the correct way. Moreover, using for example the RBF kernel $k(x,y) = exp(-\|x-y\|^2/2\sigma^2)$, is able to manage complex non-linear problems as in Figure 3c.

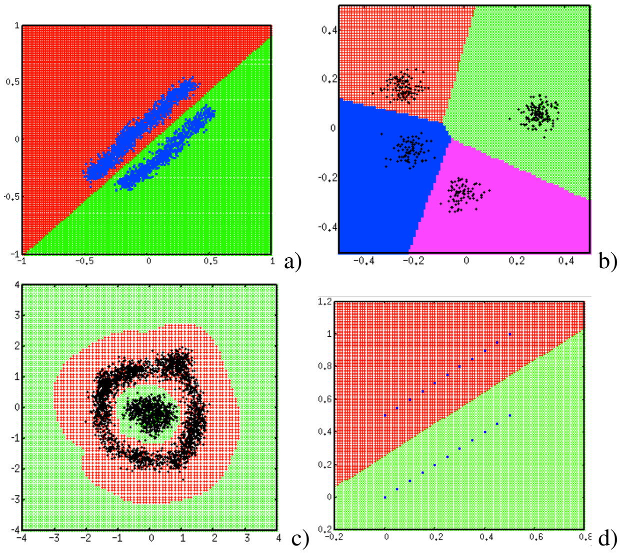

Figure 3: Some artificial problems used to verify KQC behaviour.

4 Experiments and Evaluation

In order to test quantitatively the effectiveness of the proposed quantum classifier – KQC – we set up a number of experiments, both using artificial benchmarks and real problems, and compared the KQC performances with one of the machine learning methods that usually achieve state-of-the-art performances on a large number of classification problems, that is Support Vector Machines. We relied on the SVM implementations in the SVM-light package (Joachims, 1999) and in the SVM-Multiclass package (Joachims *et al.*, 2009).

4.1 Artificial datasets

We used two artificial datasets: 2-SPIRALS and DECSIN as defined in (Segata, Blanzieri, 2009), without adding any noise to the data (see Figure 4).

They are both problems that involve a non-linear decision boundary and they are widely used for testing machine learning systems. The first dataset is composed by 628 instances and the second by 6280 instances. For both datasets $ni = no = 2$.

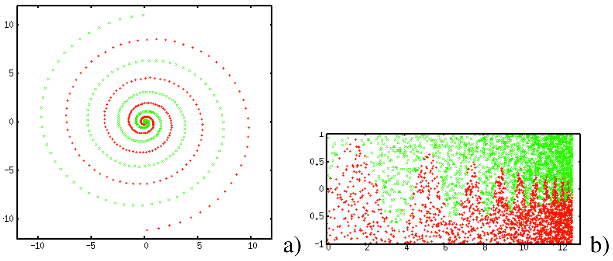

Figure 4: Artificial problems used for the evaluation. a) 2-SPIRALS, b) DECSIN.

4.2 Real problems

The two real problems used for the KQC evaluation are taken from the speech processing domain.

The first problem is a prominence identification task in connected speech (Tamburini, 2009). A subset of 382 utterances of the TIMIT Speech corpus (Garofolo *et al.*, 1990) has been manually annotated with binary prominence levels as described in (Tamburini, 2006). Extracting for each syllable the five acoustic features described in (Tamburini *et al.*, 2014), we formed a 35-feature input vector inserting the data from 3 syllables before and after the syllable. In total this dataset is composed of 4780 instance vectors.

The second problem is derived from an emotion recognition task. The E-Carini corpus (Tesser *et al.*, 2005) contains 322 utterances annotated with 7 fundamental emotions. From each utterance we extracted 1582 features using the OpenSMILE package (Eyben *et al.*, 2013) and the configuration file contained in the package for extracting the InterSpeech 2010 challenge feature set.

4.3 Results

Given the four dataset described above, we performed a number of experiments for comparing KQC with a SVM classifier. The reference metrics were precision/recall/F1 for the three binary-classified problems and the macro-averaged precision/recall/F1 for the Emotion multiclass dataset. All the experiments were performed executing a k-fold validation and optimising the classifiers parameters on a validation set. Table 1 outlines the different performances of the two classifiers when tested on the various evaluation datasets. KQC outperforms SVM in the experiments using artificial datasets and exhibit more or less the same performances of SVM on the real problems.

	KQC	**SVM**
2SPIRALS 5-fold valid.	RBF, σ=0.045 γ=0.5	RBF, σ=0.02 C=6e5

	P=1.0000
	R=0.9969
	F1=0.9984
		P=0.9532
		R=0.9776
		F1=0.9650

DECSIN 5-fold valid.	RBF, σ=0.3 γ=0.5	RBF, σ=5e-5 C=1e3

	P=0.9851
	R=0.9870
	F1=0.9860
		P=0.9827
		R=0.9805
		F1=0.9816

	KQC	**SVM**
Prominence Detection 8-fold valid.	RBF, σ=18.0 γ=0.5	LIN, C=30

	P=0.8287
	R=0.8153
	F1=0.8216
		P=0.8200
		R=0.8200
		F1=0.8200

Emotion Recognition 10-fold valid.	RBF, σ=75.0 γ=0.5	LIN, C=30

	P=0.9479
	R=0.9568
	F1=0.9523
		P=0.9793
		R=0.9728
		F1=0.9760

Table 1: KQC and SVM results (and optimal parameter sets) for the four evaluation problems.

5 Discussion and Conclusions

This paper presented a first attempt to produce a general purpose classifier based on Quantum Probability Theory. Considering the early experiments from (Liu *et al.*, 2013), KQC is more powerful and gains better performance. The results obtained on our experiments are quite encouraging and we are tempted to answer ‘yes’ to the question presented in the paper title.

This is a work in progress and the KQC is not free from problems. Despite its potential to outperform SVM using linear kernels, it is very complex to determine a tradeoff between the definition of decision boundaries with maximum margins and to maximise the classifier generalisation abilities. A long optimisation process on the training set maximise the margins between classes but could potentially lead to poor generalisations on new data. Making more experiments and evaluations in that directions is one of our future plans.

References

Blacoe W., Kashefi E. and Lapata M. 2013. A Quantum-Theoretic Approach to Distributional Semantics. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 847–857.

Busemeyer J.R. and Bruza P.D. 2012. *Quantum Models of Cognition and Decision*. Cambridge University Press.

Chen J.C.H. 2002. *Quantum Computation and Natural language Processing*. PhD thesis, University of Hamburg.

Eyben F., Weninger F., Gross F. and Schuller B. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. *ACM Multimedia (MM)*, Barcelona, 835–838.

Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett, D. and Dahlgren, N. 1990. *DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. National Institute of Standards and Technology.

Gonzàlez F.A. and Caicedo J.C. 2011. Quantum Latent Semantic Analysis. In A. Giambattista, F.Crestani (eds.),*Advances in Information Retrieval Theory*, LNCS, 6931, 52–63.

Griewank A. and Walther A. 2008. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Other Titles in Applied Mathematics 105 (2nd ed.), SIAM.

Hestenes M.R. and Stiefel E. 1952. Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49 (6), 409–436.

Ingber L. 1989. Very fast simulated re-annealing. *Mathl. Comput. Modelling*, 12 (8): 967–973.

Joachims T. 1999. Making large-Scale SVM Learning Practical. In B. Schlkopf, C. Burges, A. Smola (eds.),*Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 169–184.

Joachims T., Finley T. and Yu C-N. 2009. Cutting-Plane Training of Structural SVMs. *Machine Learning Journal*, 77 (1): 27–59.

Liu D.C. and Nocedal J. 1989. On the Limited Memory Method for Large Scale Optimization. *Mathematical Programming B*, 45 (3): 503–528.

Liu D., Yang X and Jiang M. 2013. A Novel Text Classifier Based on Quantum Computation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, 484–488.

Melucci M. and van Rijsbergen K. 2011. Quantum Mechanics and Information Retrieval. In M. Melucci and K. van Rijsbergen K (eds.),*Advanced Topics in Information Retrieval*, Springer, 33, 125–155.

Segata N. and Blanzieri E.. 2009. Empirical Assessment of Classification Accuracy of Local SVM. *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, Tilburg, 47–55.

Tamburini F. 2006. Reliable Prominence Identification in English Spontaneous Speech. *Proceedings of Speech Prosody 2006*, Dresden, PS1-9-19.

Tamburini F. 2009. Prominenza frasale e tipologia prosodica: un approccio acustico. *Linguistica e modelli tecnologici di ricerca, XL congresso internazionale di studi*, Societ di Linguistica Italiana, Vercelli, 437–455.

Tamburini F., Bertini, C., Bertinetto, P.M. 2014. Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. *Proceedings of Speech Prosody 2014*, Dublin, 285–289.

Tesser F., Cosi P., Drioli C. and Tisato G. 2005. Emotional FESTIVAL-MBROLA TTS synthesis. *Proceedings of the 9th European Conference on Speech Communication and Technology - InterSpeech2005*, Lisbon, 505–508.

Vedral V. 2007. *Introduction to Quantum Information Science*. Oxford University Press, USA.

Yeang C.H. 2010. A probabilistic graphical model of quantum systems. *Proceedings, the 9th International Conference on Machine Learning and Applications (ICMLA)*, Washington DC, 155–162.

Zuccon G., Azzopardi L.A. and van Rijsbergen K. 2009. The Quantum Probability Ranking Principle for Information Retrieval. In L.Azzopardi *et al.* (eds.),*Advances in Information Retrieval Theory*, LNCS, 5766, 232–240.