International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:8, No:5, 2014

Use of Gaussian-Euclidean Hybrid Function Based Artificial Immune System for Breast Cancer Diagnosis

Cuneyt Yucelbas, Seral Ozsen, Sule Yucelbas, Gulay Tezel

***Abstract*—**Due to the fact that there exist only a small number of complex systems in artificial immune system (AIS) that work out nonlinear problems, nonlinear AIS approaches, among the well-known solution techniques, need to be developed. Gaussian function is usually used as similarity estimation in classification problems and pattern recognition. In this study, diagnosis of breast cancer, the second type of the most widespread cancer in women, was performed with different distance calculation functions that euclidean, gaussian and gaussian-euclidean hybrid function in the clonal selection model of classical AIS on Wisconsin Breast Cancer Dataset (WBCD), which was taken from the University of California, Irvine Machine-Learning Repository. We used 3-fold cross validation method to train and test the dataset. According to the results, the maximum test classification accuracy was reported as 97.35% by using of gaussian-euclidean hybrid function for fold-3. Also, mean of test classification accuracies for all of functions were obtained as 94.78%, 94.45% and 95.31% with use of euclidean, gaussian and gaussian-euclidean, respectively. With these results, gaussian-euclidean hybrid function seems to be a potential distance calculation method, and it may be considered as an alternative distance calculation method for hard nonlinear classification problems.

***Keywords*—**Artificial Immune System, Breast Cancer Diagnosis, Euclidean Function, Gaussian Function.

I. Introduction

CANCER begins with out-of-control cleavage of one cell and results in a visible mass named tumour. Tumour can be benign or malignant. Malignant tumour grows quickly and spreads over its surrounding tissues causing their damage. Breast cancer, the second type of general cancer in women, is a malignant tissue beginning to grow in the breast. The abnormalities like presence of a breast mass, change in shape and dimension of breast, differences in the colour of breast skin, breast aches and so on are the findings of breast cancer. Cancer diagnosis is performed based on the nonmolecular criterions like tissue type, pathological properties and clinical location [1]. Worldwide, 22.9% of all cancers in women is breast. And also, 458,503 people died worldwide in 2008 due to cancer (13.7% of cancer deaths in women) [2].

The use of expert systems and different artificial intelligence techniques is increasing gradually in medical diagnosis. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. Classification systems, helping possible errors that can be done because of tired or unpractised expert to be decreased, provide medical data to be analyzed in shorter time and more comprehensive. Nowadays, new findings in cancer treatment have caused higher survival rates in cancer so in Breast Cancer. Particularly, early diagnosis can increase these survival rates in significant quantity [3]. Artificial immune systems (AISs), generated by the human immune system, are a branch of artificial intelligence (AI). Many methods have been developed so far (over 20 years) that are appropriate for different problems. Nevertheless, these methods achieved noticeable outcomes in some implementation fields such as optimization, virus detection, anomaly detection problems. But, AIS could not reach high results in classification applications yet [4]. Though the metaphors of human immune system were modeled correctly, some inadequacies like modeling linear and simple interactions may cause this failure. A significant amount of AIS methods applies the computations between the system units and inputs in the input feature space. Also, they do not include any functional or layered structure that supplies nonlinearity except for some AIS methods like [5]–[8].

As with the other clinical diagnosis problems, classification systems have been used to diagnose breast cancer illness, too. When the studies in literature about this classification application are analyzed, it can be seen that a great deal of different methods were used which achieved high classification accuracies. Among these, the authors in [9] obtained 94.74% classification accuracy using 10 fold cross validation with C4.5 decision tree method. Another study in [10] got 94.99% accuracy with RIAC method. Result of 95.06% was obtained with neuro-fuzzy techniques [11]. Authors in [12] applied three dissimilar methods to the problem which were obtained with the these results: Optimized-LVQ method's performance was 96.7%, big-LVQ method achieved 96.8% and the last method, AIRS which he submitted based on the Artificial Immune System, obtained 97.2% classification accuracy. In another study, was obtained 98.51% accuracy for breast cancer diagnosing by using Fuzzy-AIRS [3]. Reference [13] carried out a comparison between the different classifiers such as decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC).

C. Y. is with the Department of Electrical-Electronics Engineering, Selcuk University, Konya, Turkey. (phone: +90-332-2233712; fax: +90-332-2410635; e-mail: cyucelbas@selcuk.edu.tr).

S. O. is with the Department Of Electrical-Electronics Engineering, Selcuk University, Konya, Turkey. (phone: +90-332-2232035; fax: +90-332-2410635; e-mail: seral@selcuk.edu.tr).

S. Y. and G. T. are with the Department Of Computer Engineering, Selcuk University, Konya, Turkey. (phone: +90-332-2231993; fax: +90-332-2410635; e-mail: syucelbas@selcuk.edu.tr, gtezel@selcuk.edu.tr).

In this study, diagnosis of breast cancer, the second type of the most widespread cancer in women, was applied with a different distance calculation method that gaussian-euclidean hybrid function in the clonal selection models of classical AIS. Also, experimental comparison of gaussian, euclidean and gaussian-euclidean hybrid functions in AIS was implemented on Wisconsin breast cancer dataset (WBCD), which was taken from [14]. We used 3-fold cross validation method to train and test the dataset.

The other chapters of the study are organized as follows. Breast cancer diagnosis problem is explained in second section. Third section includes the natural immune system. In fourth part, AIS and used function are mentioned with explanations. The experimental results are given in fifth section, and in the last section, interpretation of results and an overall evaluation are carried out, briefly.

II. Dataset: The Winsconsin Breast Cancer

In this study, we have applied our system on the WBCD taken from UCI machine learning repository (UCI Repository of Machine Learning Databases). It consists of 683 patterns that were compiled by Dr. W.H. Wolberg at the University of Wisconsin-Madison Hospitals taken from needle aspirates from human breast cancer tissue [14]. 444 of 683 samples are benign class and the remaining 239 patterns are malignant class. The WBCD consists of nine features obtained from fine needle aspirates, each of which is ultimately represented as an integer value between 1 and 10. The features showed in Table I are numerated 1–10, with 10 being the most abnormal case. The class property was indicated as 2 for benign and 4 for malignant situations. Also, graphical representation of the eighth feature of the dataset is given in Fig. 1.

III. Natural Immune System

Natural immune system is a layered and distributed protection system versus unknown elements named as Antigen (Ag) like microbes, bacteria, viruses and so on. The system works thanks to distinguishing these reavers from the cells of body resulting with immune response. While the system consists of a great sort of elements, the most active ones in immune response are B and T lymphocytes (cells), which have receptors providing them to connect other cells including Antigens. The receptor of T cells are named TCR (T-cell receptor) while the one of the B cells are called as Antibody (Ab), the most widespread modelled unit of immune system in Artificial Immune Systems [15], [16]. General immune response to invaders is given in Fig. 2.

TABLE I
The Nine Features of Breast Cancer Dataset

Features	Value	Normalized Value
Clump Thickness	1 - 10	0 - 1
Uniformity of Cell Size	1 - 10	0 - 1
Uniformity of Cell Shape	1 - 10	0 - 1
Marginal Adhesion	1 - 10	0 - 1
Single Epithelial Cell Size	1 - 10	0 - 1
Bare Nuclei	1 - 10	0 - 1
Bland Chromatin	1 - 10	0 - 1
Normal Nucleoli	1 - 10	0 - 1
Mitoses	1 - 10	0 - 1

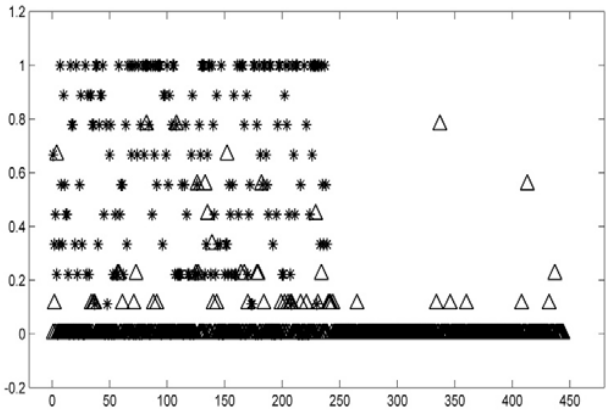

Fig. 1 Graphical representation for the eighth feature of dataset (*: malignant, Δ: benign)

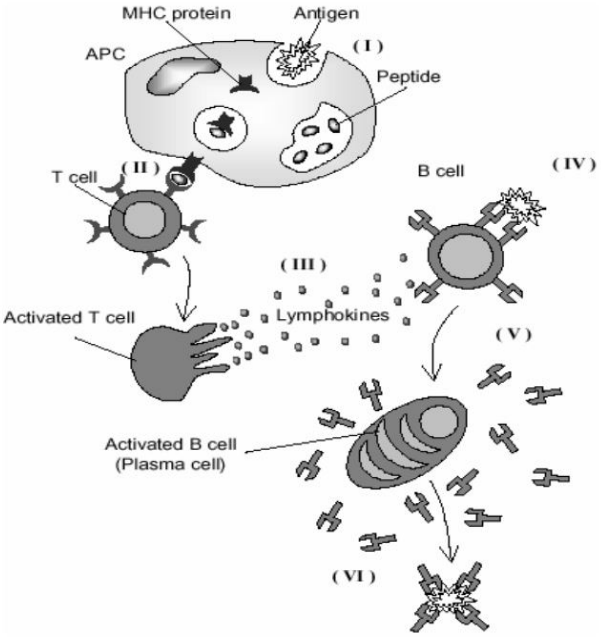

Fig. 2 General immune response to invaders [17]

As in Fig. 2, firstly, Antigen Presenting Cells (APCs) swallow and fragment Antigens into antigenic peptides (I). Secondly, the presented MHC-peptide combination on the cell

surface is known by the T-cells to be activated (II). Then, activated T cells hidden some chemicals as stimulating signals to other units in response to this recognition (III). After that B cells, one of the units that take these signals from the T cells actuated the recognition of Antigen by their Antibodies occurred in the same time (IV). Afterwards, B cells transmute plasma cells (V). And finally, secreted Antibodies keep in the existing Antigens and passivate them signalling other items of immune system to demolish the Antigen-Antibody complex (VI) [16], [17].

IV. Artificial Immune System and Used Functions

Artificial Immune System is a resource restricted supervised learning algorithm inspired from natural immune system. In this algorithm, the used immune systems are resource competition, clonal selection, affinity maturation and memory cell creation. The feature vectors presented for training and test operations are called as Antigens while the system units are named as B cells. Firstly, all of training antigens are presented and then the memory cells are composed. Afterwards, the memory cells are used to classify test antigens. This algorithm consists of two stages that are training and test.

In training process, clonal selection mechanism in the immune system was modeled and used euclidean, gaussian and gaussian-euclidean functions for calculation of distance between Ag and Ab. Flowchart of the performed system is given in Fig. 3. Input data and system units are represented as Antigens (Ags) and Antibodies (Abs), respectively. For each submitted Ag, an Ab is generated that gets to know the presented Ag, and the class of this Ab gives the class information of the presented Ag. In our system Shape–space representation was used. The input data represented as Ags are modeled as vectors with L features:

$$Ag = [Ag_{x1}, Ag_{x2}, ..., Ag_{xL}]^T$$

Abs (or system units) are also represented as L-dimensional vectors:

$$Ab = [Ab_{y1}, Ab_{y2}, ..., Ab_{yL}]^T$$

We use Euclidean, Gaussian and Gaussian-Euclidean functions to calculate the similarities between Ags and Abs. The similarities that distances between Ab and Ag are calculated. The result which is close to 1 stands for maximum of similarity between Ab and Ag. Then, a certain number (K) of Abs, whose similarities are maximum, are selected for cloning and mutation. After that, these selected Abs are subjected cloning and mutation processes according to their similarity values.

Euclidean distance is used for both affinity and stimulation value calculations as in (1):

$$Euclidean\ Dist. = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

where x and y refer feature vectors while n is the number of attributes in data.

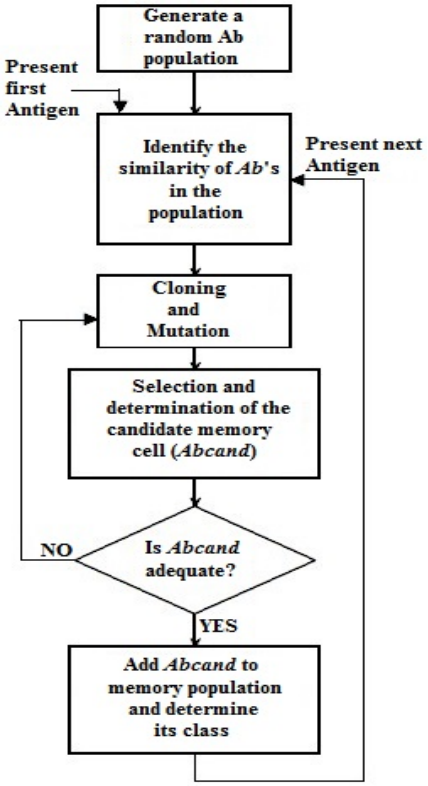

Fig. 3 Flowchart of the performed system

The Gaussian distribution is also mostly called the "normal distribution" and is usually defined as a "bell-shaped curve". The Gaussian distribution is a continuous function which approaches the certain binomial distribution of samples. Normal distributions are immensely significant in statistics and are frequently used in the natural and social sciences for real-valued random samples whose distributions are not identified [18].

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

The parameter μ in (2) is the mean or expectation of the distribution. The parameter σ is standard deviation of the samples, and σ^2 is their variance. x is related unit in this equation.

As in Fig. 4, interaction between Ab and Ag in Gaussian distribution is given. According to this figure, the similarity calculations between Ab and Ag are performed according to (2). If the similarity between Ab and Ag approaches the maximum, the affinity value closes to 1, too. In this way,

distance calculation between units is performed and finally memory antibodies are occured by the obtained values.

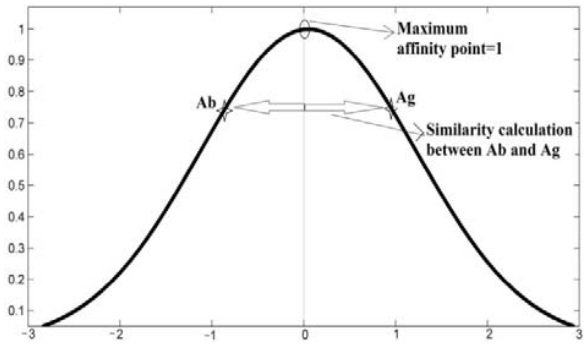

Fig. 4 Gaussian distribution in this study

In this study, both Gaussian and euclidean functions are separately used for distance computation between Ab and Ag in the Gaussian-Euclidean calculation method. Then, the best result of these is taken and the related Ab is cloned and mutated to exceed the threshold level. In this manner it is intended to obtain a higher classification rate and has been successful. The related application results are presented in 5th section.

In test process, the training part gives memory Abs and the class information of these Abs. In this phase, these outputs are used to determine the classes of test Ags as follows:

1) For each test Ag (Ag_t , t:1, 2,....):
- calculate similarity of memory Abs to present Ag_t;
- identify the class of memory Ab whose similarity is maximum to Ag_t as the class of $Ag_t \rightarrow s\_test(t)$.

And thereby, the class information of every Ag is determined as *s_test(t)*. Classification accuracy is calculated as follows:

$$Classification\_accuracy = \sum_{i=1}^{N_t} \frac{true\_s(i)}{N_t} \qquad (3)$$

$$true\_s(i) = \begin{cases} 1, & s\_test(i) = s(i) \\ 0, & otherwise \end{cases}$$

Here, *s_test(i)* is the estimated class of Ag_i by the system. N_t is the total number of the test data, and *true_s(i)* is the original class of Ag_i.

V. The Experimental Results

In this section, we present performance results of the used functions that are euclidean, gaussian and gaussian-euclidean for Wisconsin breast cancer dataset.

In the training phase of the functions based AIS, there are a few user-dependent parameters that affect the performance. Related parameters are as follows:

Sc (stopping criterion): threshold value that is used to identify whether the candidate memory cells will be added to the memory population or not. We defined this parameter as 0.88 value.

itnum: The parameter defines the maximum number of iterations. Therefore, it should be chosen high enough to determine candidate memory cells. We identified it as 100 iterations.

M: Number of Abs in Ab_pop. We selected 10, 30, 50, 70 and 100, respectively.

K: the number of B cells chosen for cloning and mutation. It was taken as M/3. For instance, if M is 30, K will be 10 (M/3=10).

Also, we used 3-fold cross validation method to train and test the dataset as seen in Table II.

TABLE II
3-Fold Cross Validation for WBCD

Fold-k		Fold-1	Fold-2	Fold-3
Training dataset	Malignant	137	154	187
	Benign	319	302	267
	Total	456	456	454
Testing dataset	Malignant	102	85	52
	Benign	125	142	177
	Total	227	227	229
Overall Total		683	683	683

Equation (3) in Section IV is used to measure the classification accuracies for the dataset. The experimental results that are obtained classification accuracies are shown in Table III. As shown in this table, the highest classification accuracy was obtained by Gaussian-Euclidean function as 95.31% with 16 memory cells. The other best results were listed as %94.78 with 21 memory cells and 94.45% with 139 memory cells for euclidean and gaussian functions, respectively. The results are also presented in Fig. 5 graphically.

TABLE III
Classification Accuracies for Wisconsin Breast Cancer Dataset

Functions	M	Classification accuracy (%)	Number of memory cells (Mem_pop)
Euclidean	**10**	**94.78**	**21**
	30	94.13	17
	50	94.23	16
	70	94.09	16
	100	94.38	15
Gaussian	10	93.61	141
	30	93.03	133
	50	94.10	141
	70	94.12	138
	100	**94.45**	**139**
Gaussian and Euclidean	10	94.79	20
	30	94.63	17
	50	94.98	17
	70	95.01	16
	100	**95.31**	**16**

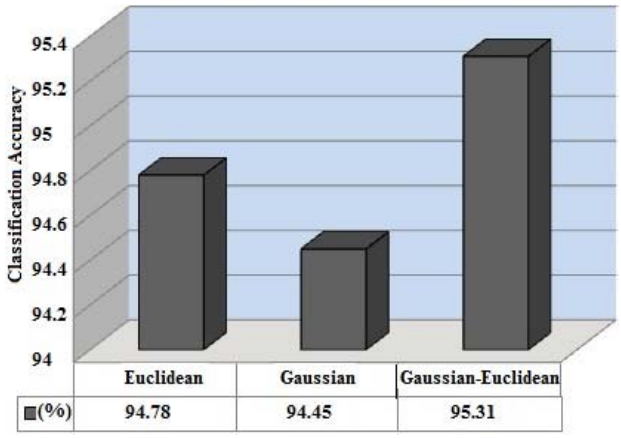

Fig. 5 The effects of three functions on classification accuracies for WBCD dataset

As seen in Fig. 5, the classification accuracies for three functions, Gaussian-Euclidean hybrid function method was obtained as the most efficient method. Euclidean and Gaussian functions followed this method.

VI. CONCLUSION

In that study, different distance calculation functions which are euclidean, gaussian and gaussian-euclidean were experimented in the standard clonal selection model of classical AIS.

In the application phase of this study, Wisconsin breast cancer dataset (WBCD), which was taken from the University of California, Irvine Machine-Learning Repository, was used for training and testing steps. In the classification of the illness, the experiments were performed to see the effects of the new distance calculation method. The highest results were obtained as 94.78%, 94.45% and 95.31% for euclidean, gaussian and euclidean-gaussian functions, respectively. According to the application results, gaussian-euclidean hybrid function based AIS demonstrated an important high performance with the minimum number of antibodies as 16 memory cells.

We can say that if the parameters such as itnum, M, K and especially Sc are arranged, better results can be obtained. We believe that higher performance results can be reached for Sc's value closer to 1. Also, the results taken in this study can be considered good according to the outcomes of similar studies in the literature.

We suggest that euclidean-gaussian function based AIS can be applied the different types of data, too.

ACKNOWLEDGMENT

This study is supported by the Scientific Research Projects of Selcuk University.

REFERENCES

[1] X.L. Du, C.R. Key, C. Osborne, J.D. Mahnken and J.S. Goodwin, "Discrepancy between Consensus Recommendations and Actual Community Use of Adjuvant Chemotherapy in Women with Breast Cancer," *Annals of Internal Medicine*, vol. 138, 2003, pp. 90-97.

[2] Wikipedia, Breast Cancer. Available at: *http://en.wikipedia.org/wiki/Breast_cancer*, (last accessed: 25 December 2013).

[3] P. Kemal, S. Seral, K. Halife and G. Salih, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," *Expert Systems with Applications*, vol. 32, 2007, pp. 172–183.

[4] O. Seral, G. Salih, K. Sadık and L. Fatma, "Use of Kernel Functions in Artificial Immune Systems for the Nonlinear Classification Problems," *IEEE Transactions On Information Technology In Biomedicine*, vol. 13, no. 4, July 2009, p. 621.

[5] J.H. Carter, "The immune system as a model for pattern recognition and classification," *J. Amer. Med. Inf. Assoc.*, vol. 7, no. 1, 2000, pp. 28–41.

[6] W.-D. Sun, Z. Tang, H. Tamura, and M. Ishii, "A hierarchical artificial immune architecture and its applications," *in Proc. SICE Annu. Conf.*, Fukui, Japan, 2003, pp. 3265–3270.

[7] D. Dasgupta, S. Yu, and N. S. Majumdar, "MILA-multilevel immune learning algorithm," *(Lecture Notes in Computer Science) in Proc. GECCO 2003*, vol. 2723, 2003, pp. 183–194.

[8] P. Bentley, J. Greensmith, and S. Ujjin, "Two ways to grow tissue for artificial immune systems," *(Lecture Notes in Computer Science) in Proc. ICARIS 2005*, vol. 3627, 2005, pp. 139–152.

[9] J.R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, 1996, pp. 77-90.

[10] H.J. Hamilton, N. Shan and N. Cercone, "RIAC: A Rule Induction Algorithm Based on Approximate Classification," *Tech. Rep. CS 96-06*, Regina University, 1996.

[11] D. Nauck, and R. Kruse, "Obtaining Interpretable Fuzzy Classification Rules From Medical Data," *Artificial Intelligence in Medicine*, vol. 16, 1999, pp. 149-169.

[12] E.D. Goodman, C.L. Boggess and A. Watkins, "Artificial Immune System Classification of Multiple-Class Problems, In Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming," *Complex Systems and Artificial Life*, vol. 12, 2002, pp. 179-184.

[13] I. Salama, M. B. Abdelhalim, and Z.M.A. Gouda, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis," *Seventh International Conference on Compute Engineering & Systems (ICCES)*, 21-29 Nov. 2012, pp. 180-185.

[14] C.L. Blake and C.J. Merz, UJI Repository of Machine Learning Databases, Available at: "*ftp://ftp.ics.uci.edu/pub/machine-learning-databases*" (last accessed: 7 April 2010).

[15] O.F. Nasaroui, F. Gonzalez and D. Dasgupta, "The Fuzzy Artificial Immune System: Motivation, Basic Concepts, and Application to Clustering and Web Profiling," *International Joint Conference on Fuzzy Systems*, 2002, pp. 711-717.

[16] S. Şahan, H. Kodaz, S. Güneş and K. Polat, "A New Classifier Based on Attribute Weighted Artificial Immune System," *Lecture Notes in Computer Science (LNSC 3280)*, 2004, pp. 11-20.

[17] L.N. De Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer-Verlag Press, 2002.

[18] Wikipedia, Normal distrubition. Available at: "*http://en.wikipedia.org/wiki/Normal_distribution*" (last accessed: 25 December 2013).