Text Processing Using Support Vector Machine

P.Samba sivarao[1] , CH.Prasannalaksmi[2] ,G.Sowmyasree[3] , J.Ramyakeerthi[4],M.Jayatarunl[5]

[2,3,4,5] *B.tech Computer Science and Engineering, Lendi Institute of Engineering and Technology,*
Jonnada, Vizianagaram, AP, India.

[1] *Assistant Professor, Department of Computer science and Engineering, Lendi Institute of Engineering and Technology*
, *Jonnada, Vizianagaram, AP, India.*

[1]Kbsr1@gmail.com
Prasannachoubay gmail.com
[3]gopisowmyasree@gmail.com
[4]ramyakeerthi537@gmail.com
[5]tharuni2116@gmail.com

Abstract— **In this present trend Authentic plays a very crucial role.our project is a blend of SVM. SVM could produce pool based active learning.Our theme is to access to a pool of unlabeled instances and can request the labels for some number of them. In the above instance, if the classification categories are not approximately equally represented. We propose an efficient method based on active learning strategy to retrieve large text categories. At each feedback step, the system optimizes the text presented to the user in order to speed up the retrieval. Here we use SVM method to make our project different from normal text Processing. Support vector machines (SVM) have met with significant success in numerous real-world learning tasks. However, like most machine learning algorithms, they are generally applied using a randomly selected training set classified in advance. We introduce a new algorithm for performing active learning with support vector machines, i.e., an algorithm for choosing which instances to request next. We provide a theoretical motivation for the algorithm using the notion of a version space our active learning method can significantly reduce the need for labelled training instances in both the standard inductive and transductive settings. One of the most critical problems for machine learning methods is over fitting. The over fitting problem is a phenomenon in which the accuracy of the model on unseen data is poor where as the training accuracy is nearly perfect. This problem is particularly severe in complex models that have a large set of parameters. In this paper, we propose a deep learning neutral network model that adapts the support vector data description (SVDD). The SVDD is a variant of the SVM, which has high generalization performance by acquiring a maximal margin in one class classification problems. The proposed model strives to obtain the representational power of deep learning. Generalization performance is maintained using the SVDD. The experimental results showed that the proposed model can learn multiclass data without severe over fitting problems**

Keywords— **Active learning, Support vector machine, Classification, Selective sampling, Support vector data description, Deep learning, Pattern recognition, Generalization, Relevance feedback.**

Introduction

Machine learning algorithms for pattern recognition inherently rely on the characteristics of extracted features. Deep learning models learn complicated functions with large data sets to extract high-level features automatically through deep-layered neural network structures Conventional deep neural network architectures use multiple hidden layers instead of a single hidden layer. However, it is usually difficult to identify an adequate learning algorithm to train those inter connection weights with multiple hidden layers. Thus, the interconnection weights in multiple layers are expected to replace manual domain-specific feature engineering in the case of conventional machine learning.

Moreover, recent neuroscience research has provided further elucidation and background information for efficiently constructing deep feature extraction the SVM is a supervised machine learning algorithm. An SVM with a shallow layer structure is commonly used for classification and regression, particularly with the kernel trick, which performs predictions for new inputs depending only on the kernel function evaluate data sparse subset of training data points it constructs an optimal decision hyperplane in the context of the maximal margin. In the SVM, maximal margin guarantees a high generalization performance SVMs are usually sensitive to noise patterns or outliers because a relatively small number of mislabeled examples or outliers can dramatically decrease the performance. In other words, an outlier can critically affect the decision boundary and calculation of the margin.

TEXT PROCESSING:

In computing, the term text processing refers to the discipline of mechanizing the creation or manipulation of electronic text. Text usually refers to all the alphanumeric characters specified on the keyboard of the person performing the mechanization, but in general text here means the abstraction layer that is one layer above the standard character encoding of the target text. The term processing refers to automated (or mechanized) processing, as opposed to the same manipulation done manually.

Text processing involves computer commands which invoke content, content changes, and cursor movement, for example to search and replace format generate a processed report of the content of, or filter a file or report of a text file.

The text processing of a regular expression is a virtual editing machine, having a primitive programming language that has named registers (identifiers), and named positions in the sequence of characters comprising the text. Using these "text processor" can, for example, mark a region of text, and then move it.

The text processing of a utility is a filter program, or filter. These two mechanisms comprise text processing.

SUPPORT VECTOR DATA DESCRIPTION:

The SVDD is a variant of the conventional SVM, as mentioned previously; therefore, both the SVM and the SVDD are briefly described in this subsection. The SVM is a supervised learning method that is widely used in classification and regression tasks. For the linearly separable problem, the SVM obtains the maximal margin hyperplane, and the distance from the hyperplanet the nearest data point one side is maximized. The SVM support vectors indicate the training data points closest to the hyperplane. To define decision boundaries, the SVM formulation therefore depends on the support vectors, which are the sample points at the discriminatory boundary and within the separating margins of the two classes. The SVDD has been successfully applied in a wide variety of application domains, such as handwritten digit recognition, facial recognition, and anomaly detection. The closed hypersphere of SVDD separates the inner region with a high density of data from the outer region with low density. In general, because the probability density estimation of the target data requires numerous samples, the SVDD creates a description of training data by constructing the sphere around the given data; the boundary can be used to detect which objects resemble the learned training set.

Fig. 1: shows the confidence data number of the Climates data set according to the depth of the model. The bar heights mark the number of confidence data; the white and black areas represent the portions of positive and negative classes, respectively .As the depth of the model increases by learning in a greedy manner, the confidence data for the model increases.

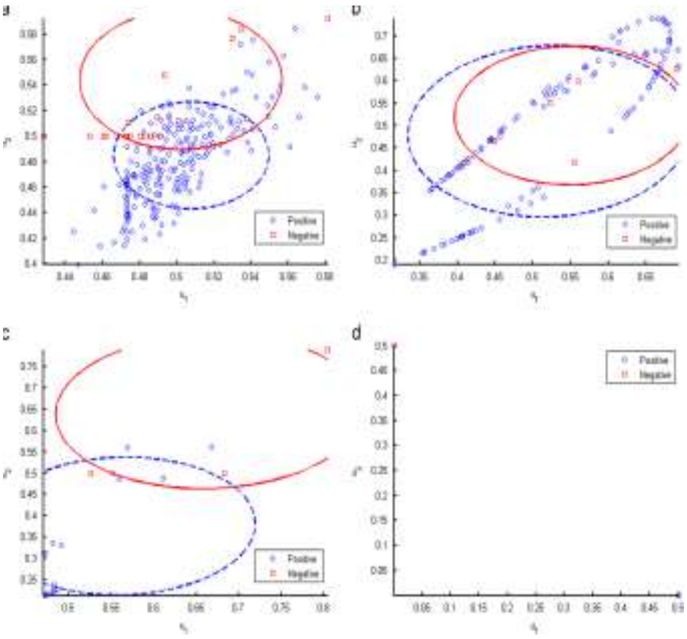

Fig 1:decision boundaries of deep layers (a)Second layer (b)Eighth layer (c)fourteenth layer (d)Twetieth layer

PROPOSED MODEL:

In the proposed model, as in conventional artificial neural networks, the input vector of one hidden layer is the concatenation of output values of the previous layer .As a result, the dimensionality of the input vector for the hidden layers is fixed to k for k-class problems. However, in the proposed model, unlike conventional artificial neural networks, both the final layer and intermediate layer scan provide their decision on the test data. Because the description can be viewed as the value related to a confidence measure on the decision of each SVDD node, if the data do not satisfy a confidence rule ,the given data are re-entered into the next layer with the transformed feature vector. Otherwise, the current layer presents a final decision on the data. This is illustrated by the two-dimensional example in Fig. 2. In Fig. 2, a1 is the center of a class sphere, and u1 and u2 represent each axis of the two-dimensional data. The solid line depicts the boundary of the sphere the area of the sphere that contains a test sample is identified as the class label. The dashed line denotes the class confidence region. If a given data sum x1 with two dimensions lies in the confidence region, it is determined to be the class label. If it lies outside the confidence region, the current SVDD node does not make a decision for this datum. In the layer consisting of multiple SVDD nodes, if no SVDD nodes make a decision or several nodes classify the datum as their class simultaneously, the datum is considered as suspicious one,

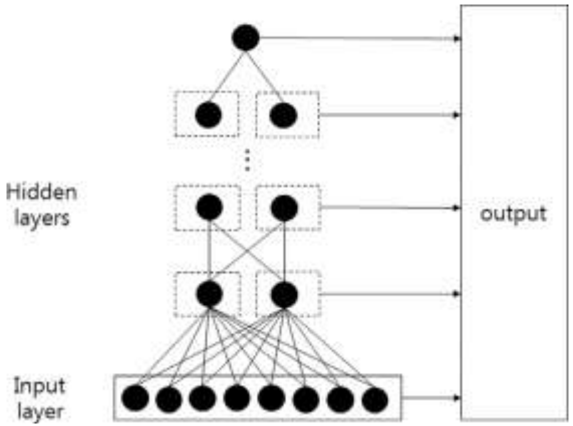

Fig 2:Proposed Model

References

[1] I. Arel, D.C.Rose, T.P.Karnowski, Deep machine learning-a new frontier in artificial intelligence research [research frontier], Comput.Intell.Mag.IEEE5 (2010)13–18.

[2] Y. Bengio, Learning deep architectures for AI, Found. Trends sMach. Learn.2 (2009)1–127.

[3] G. Tesauro , Practical issues in temporal difference learning, in: Reinforcement Learning, Springer,1992,pp.33–53.

[4] P.E.Utgoff,D.J.Stracuzzi,Many layered learning, Neural Comput.14(2002) 2497–2529.

[5] K. Fukushima, Neo cognitron :a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern . 36 (1980)193–202.

[6]D.H.Ackley, G.E.Hinton, T.J.Sejnowski, A learning algorithm for Boltzmann machines, Cogn.Sci.9(1985)147–169.

[7] P.Smolensky, Information Processing in Dynamical Systems:Foundations of HarmonyTheory,1986.

[8] G.E. Hinton, S.Osindero, Y.-W.Teh, A fast learning algorithm for deep belief nets, NeuralComput.18(2006)1527–1554.

[9] N. LeRoux, Y.Bengio, Representational power of restricted Boltzmann machines anddeep belief networks, Neural Comput. 20(2008)1631–1649.

[10]P. Vincent, H. Larochelle, I. Laja: e, Y.Benjio,P..A. Manzag, stacked denoising auto encoders: learning usefull representations

[11] G. Hinton, Apractical guide to training restricted Boltzmann machines, Momentum 9(2010)926.

[12] Y.Lecun, Y.Bengio, Convolutional Network for images, speech and Time series.in: The Handbook of Brain Theory and Neural Network,3361,1995.

[13] G.H. Golub, M.Heath, G.Wahba, Generalized cross-validation as a method or choosing a good ridge parameter, Technometrics21(1979)215–223.

[14]G.E.Hinton,N.Srivastava, A.Krizhevsky, I.Sutskever, R.R. Salakhut dinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint, arXiv:1207.0580, 2012.

[15]I.J.Goodfellow,D.Warde Farley, M.Mirza, A.Courville, Y.Bengio, Maxout networks, arXiv preprint, arXiv:1302.4389, 2013.

[16]L.Wan,M.Zeiler,S.Zhang,Y.L.Cun,R.Fergus,Regularization of neural networks using drop connect, in: Proceedings of the30thInternational Conference on Machine Learning(ICML-13),2013,pp.1058–1066.

[17] C. Cortes,V. Vapnik, Support-vector networks, Mach. Learn. 20(1995) 273–297.

[18]C.M.Bishop,Pattern Recognition and Machine Learning, Springer, NewYork, 2006.

[19] J. ShaweTaylor,P.L. Bartlett, R.C.Williamson, M.Anthony, Structuralrisk minimization overdata-dependenthierarchies, Inf.TheoryIEEETrans.44 (1998)1926–1940.

[20] D.M. Tax, R.P.Duin, Support vector data description, Mach.Learn.54(2004)