تعلم البيانات المهيكلة ببيان باستخدام الشبكات العصبية

عبدالله محسن محمد الجفري

إشراف
د. محمد معين الدين الانصاري
ا.د. عبيد السقاف

المستخلص

ازدادت في الحقبة الأخيرة كمية البيانات المسجلة من مختلف نواحي الحياة. في أغلب الحالات تكون هذه البيانات في حالة خامة. للاستفادة من البيانات المجموعة، يعمد المختصون في المجال إلى الاستعانة بخوارزميات خاصة لمعالجة هذه البيانات. أدى التطور التقني في مجال الحواسب في الفترة الأخيرة إلى تصدر خوارزميات الشبكات العصبية. ينقسم عمل الشبكات العصبية إلى مرحلتين. في المرحلة الأولى تقوم الشبكة بالتعلم من البيانات، وفي المرحلة الثانية يمكن الاستفادة من الشبكة للقيام بتوقعات واتخاذ قرارات بناء على البيانات المجموعة. هناك طرق عديدة لتصميم شبكة عصبية، وقد تم تصميم العديد من الشبكات المختصة بحسب نوع البيانات. تم التوجه في هذه الرسالة إلى البيانات المهيكلة ببيان (مخطط). يمكن تمثيل هذه البيانات برسم بياني تجريدي، حيث تمثل كل نقطة من البيانات برأس (أو عقدة) وبعد ذلك يتم ربط كل الرؤوس ذات الصلة بضلع (خط). تتواجد المعلومات في البيانات ذات البيان في النقاط البيانية نفسها وفي الهيكل الذي يربط بعض النقاط ببعضها. استخدام الشبكات العصبية التقليدية مع هذه البيانات يؤدي إلى فقد المعلومات المتضمنة في الهيكل التركيبي للبيان. هناك عدد من الشبكات العصبية التي صممت خصيصا لهذا النوع من البيانات بنتائج متفاوتة. اعتمد في تصميم الشبكة العصبية البيانية الجديدة في هذه الأطروحة على خوارزميات الانتشار في البيانات ذات البيان. تقوم خوارزميات الانتشار بدمج المعلومات الموجودة في النقاط البيانية مع التركيب الهيكلي الذي يربط النقاط ببعضها البعض. هناك نوعان من خوارزميات الانتشار المسخدمة هنا: النوع الأول ناتج من مصفوفة لابلاس للبيان (المخطط)، والنوع الثاني ناتج من عملية السير العشوائي. تم اقتراح نموذج انتشار جديد قائم على دمج النماذج المعتمدة على مصفوفة لابلاس والسير العشوائي. تمت تسمية الشبكة العصبية الجديدة بـ"شبكة التضمين بالانتشار البياني المركب". كي تتم مقارنة الشبكة المقترحة مع التصاميم السابقة بشكل عادل، تم الاعتماد على مجموعتين من شبكات الاستشهاد القياسية: Cora وCiteseer. هذه المجموعتين مستخدمتين بكثرة في البحث العلمي لاختبار نماذج التعلم على البيانات ذات البيان. حققت الشبكة الجديدة نسبة توقع صحيحة ٨٤,٩ % في شبكة Cora، ونسبة توقع صحيحة ٧٣,٤ % في شبكة Citeseer. نتيجة شبكة Cora تمثل تحسن بـ 1.9 % عن أفضل نتيجة سابقة، وتحسن بـ 0.5 % في شبكة Citeseer. تثبت هذه النتيجتان تفوق الشبكة العصبية الجديدة في تعلم البيانات ذات البيان والقيام بالتوقعات فيها.

Graph-Structured Data Learning Using Neural Networks

By
Abdullah Mohsen Al-Gafri

Supervised By
Dr. Muhammad Moinuddin
Prof. Ubaid M. Al. Saggaf

ABSTRACT

The technological advancements of our age are accompanied by large amounts of varying data. Data taken from various aspects of life and physical phenomena are considered raw. Raw data needs to be processed in order to deduce useful information that facilitates the decision making process. Data pertaining to different phenomena come in varying types. The way data is mathematically described affects the choice of the data processing tools. Machine learning algorithms and especially artificial neural networks have proved to be excellent data processing tools. They enable the user to learn from the given data and make predictions that would traditionally require a human operator to make. The literature is rich with endeavors to design neural network models with high prediction accuracy for many real life challenges. Generic neural network models can be used for any type of data with varying degrees of success. Research in neural network models strive to fully understand the properties of the data at hand. This understanding leads to designing specialized learning models that can exploit the given data and process it most efficiently. For example, Recurrent Neural Networks (RNNs) are known to be suitable for time-domain data, and Convolutional Neural Networks (CNNs) are excellent for learning from image data. The graph domain is an abstract mathematical representation that can represent many real life data. Other types of data such as time series, 2D images and 3D meshes can also be naturally adapted to the graph domain. Therefore, there is a need to develop specialized neural network models capable of learning graph structured data as efficiently as possible.

The aim of this thesis is to design a neural network model that is able to learn

from graph data and make predictions with relatively high accuracy. This work focuses on semi-supervised learning of graph data. This designation entails the availability of some labeled data point in the learning phase. The model should be able to predict the labels of the data points whose labels were not visible in the training phase. The thesis provides an overview of five prominent neural network models. Namely the models are: Graph Neural Network (GNN), Graph Convolutional Networks (GCNs), Network of GCNs (N-GCN), Graph Attention Networks (GATs), and the Attention-Based Graph Neural Networks (AGNNs). The thesis then details three graph diffusion processes. Diffusion processes are used to propagate the data through the underlying graph structure. The first two diffusion processes are formulized by two minimization problems. The two minimization problems aim to fulfill the smoothness assumption. This assumption entails the expectation that graph nodes close to each other should exhibit similar features. The smoothness assumption is also balanced by a term that restrains the data points from deviating too much from their original values after the diffusion process. The third diffusion process is based the idea of graph random walk. We propose using a combination of the three diffusion processes as a new process. The new combined diffusion process is used in designing a neural network model for graph structured data. The proposed model is named Combined Graph Diffusion Embedding Network (CGDEN). The model was tested on two benchmarking citation datasets; Cora and Citeseer. The classifying accuracy of the model was compared against a number of baselines from the literature. The model achieved a correct classifying accuracy of 84.9% on the Cora dataset, and 73.4% on the Citeseer dataset. The Cora result is 1.9% higher than the previous best performance, while the Citeseer result had a 1.3% improvement on the previous best result. The state-of-the-art results of the model on the standard benchmarking dataset prove the ability of the model to efficiently learn from and make predictions on graph structured data. The citation datasets are modeled as undirected and unweighted graphs. For future works; the CGDEN model could be modified to accept directed and weighted graphs as inputs. There is also a need to find a systematic way to find the optimum values for the hyper-parameters α and β in the model.