

**دانشگاه صنعتی شیراز**

**دانشکده مهندسی کامپیوتر و فناوری اطلاعات گروه نرم‌افزار**

**گزارش سمینار کارشناسی‌ارشد**

**در رشته مهندسی کامپیوتر گرایش شبکه**

بررسی روش های استخراج معنی از داده های شبکه های اینترنت اشیا با استفاده از روش های یادگیری ماشین روی داده های بزرگ

**دانشجو:**

**سینا عطایی**

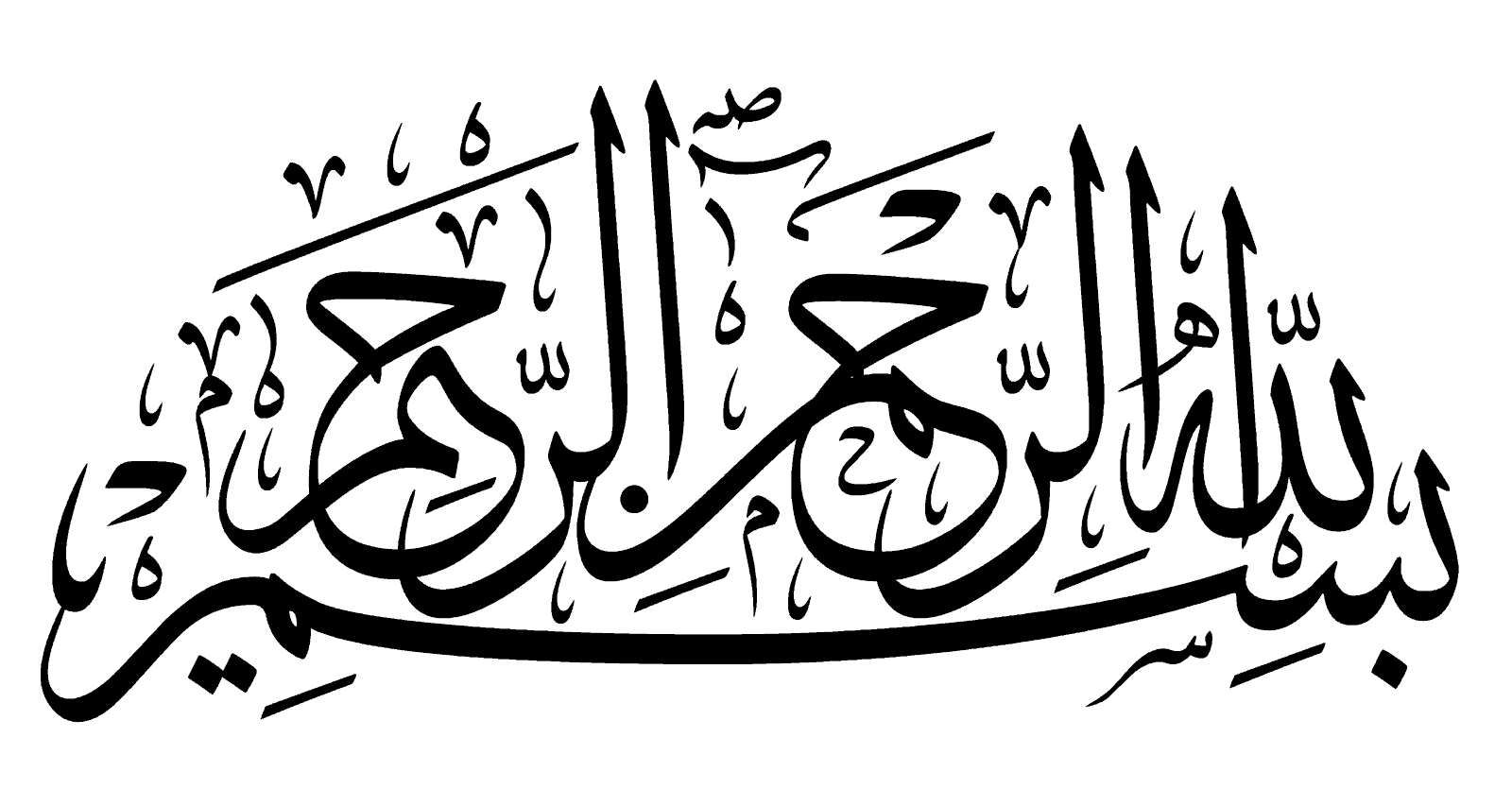
استاد راهنما:

دکتر جاویدان

استاد مشاور:

دکتر سرور

بهمن 97



چکیده

بررسی روش های استخراج معنی از داده های شبکه های اینترنت اشیا با استفاده از روش های یادگیری ماشین روی داده های بزرگ

نگارش :

از داده کاوی به عنوان یک ابزار قدرتمند در استخراج دانش از داده ها استفاده می شود. با توجه اینکه حجم وسیعی از اطلاعات در حال تولید شدن می باشد به داده کاوی برای تحلیل داده ها و کشف دانش موجود در آن ها نیاز خواهیم داشت. داده کاوی در زمینه های مختلفی از جمله پزشکی، مهندسی و اینترنت اشیاء کاربرد دارد. داده کاوی اینترنت اشیاء در سال های اخیر مورد توجه زیادی قرار گرفته است. در ابتدا مفاهیم مربوط به یادگیری ماشین، داده کاوی روی داده های بزرگ و کاربرد آن در حوزه اینترنت اشیاء را بررسی می کنیم و سپس برخی تحقیقات انجام شده در داده کاوی اینترنت اشیاء در سال های اخیر را مورد بررسی قرار می دهیم.

**کلمات کلیدی :** هوش مصنوعی، یادگیری ماشین، داده کاوی، اینترنت اشیاء، داده های بزرگ

فهرست مطالب

[فصل دوم: مروري بر ادبیات تحقیق 7](#_Toc536213813)

[1-1- مقدمه 8](#_Toc536213814)

[1-2- تعاریف و ادبیات پایه مربوط به داده کاوی 8](#_Toc536213815)

[1-3- چگونگی برخورد با داده های حجم بالا 8](#_Toc536213816)

[1-3-1- k نزدیک ترین همسایه 9](#_Toc536213817)

[1-4- نتیجه گیری 10](#_Toc536213818)

[فصل سوم: مروری بر تحقیقات اخیر انجام شده 11](#_Toc536213819)

[3-1- مقدمه 12](#_Toc536213820)

[3-2- مروری بر ادبیات موضوع 12](#_Toc536213821)

[3-3- نتیجه گیری 21](#_Toc536213822)

[مراجع 22](#_Toc536213823)

فهرست شکل‌ها

[شکل 1-1 دسته بندی روش های خوشه بندی 29](#ش4)

فهرست کلمات اختصاری

ANFIS Adaptive Neuro Fuzzy Inference System

SEER the Surveillance, Epidemiology, and End Results

SVM Support Vector Machine

aiNet-K Artificial Immune Network + k-means

ANOVA ANalysis Of VAriance

UCI University of California

M-CLUBS Microarray data CLustering Using Binary Splitting

KEEL Knowledge Extraction based on Evolutionary Learning

MTD-KNN Mega Trend Diffusion – K-Nearest Neighbor

MTD-SVM Mega Trend Diffusion – Support Vector Machine

CPAR Classification Based on Predictive Association Rule

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies

DENCLUE Clustering Based on Density Distribution Functions "DENsity-based CLUstEring

C&RT Classification and Regression Trees

CHAID CHi-squared Automatic Interaction Detection

NCBI National Center for Biotechnical Information

GEO Gene Expression Omnibus

DNF Disjunctive Normal Form

EM Expectation Maximization

ROC Receiver operating characteristic

Opt-aiNET An Immune Network Algorithm for Optimisation

PSO Particle swarm optimization

OptiGrid Optimal Grid

FCM Fuzzy C-Means

فصل دوم: مروري بر ادبیات تحقیق

* 1. مقدمه

در این فصل ابتدا به تعریف داده کاوی پرداخته می­شود. سپس به صورت خاص‌تر، تعریف داده کاوی در IoT بیان می‌گردد.

* 1. تعاریف و ادبیات پایه مربوط به داده کاوی

داده کاوی[[1]](#footnote-1)چیست؟

در لغت به معنی بررسی و کاوش داده ها می باشد.

از لحاظ مفهومی داده کاوی عبارت است از:

* ترکیب روش های تحلیل داده ها و الگوریتم های پیشرفته برای پردازش داده های با حجم بالا می باشد. داده کاوی امکان تحلیل نوع داده های جدید و همچنین روش های جدیدی برای تحلیل نوع داده های قدیمی به ما می دهد ]1[.
* فرآیندی است که به کمک کامپیوتر برای کاوش و تحلیل مجموعه های عظیم داده ها انجام می شود تا مفهموم موجود در داده ها را استخراج کند [2].
* داده کاوی فرایند استخراج داده های مفید یا الگو از داده خام می باشد ]6[.

فرض کنید یک مقدار مشخص از داده داریم، و به دنبال یک حالت مشخص از یک رویداد درون این داده ها می باشیم. می شود انتظار داشت که این رویداد بخصوص رخ دهد، حتی اگر داده ها کاملا تصادفی باشند. تعداد رویداد ها با افزایش اندازه داده ها، افزایش می یابد. اما این روی داد ها فریبنده هستند؛ زیرا داده های تصادفی همیشه تعدادی خصوصیت غیر عادی دارند که به نظر مهم می آیند اما اینطور نیست. یک تئوری آماری به نام تصحیح بنفرونی1، یک روش آماری ارائه می دهد که می تواند اکثر رویداد های فریبنده را تشخیص دهد. بدون ورود به جزئیات آماری، ما یک روش غیر رسمی بر اساس اصل بنفرونی ارائه می کنیم، که کمک می کند به اشتباه رویداد های تصادفی را پردازش نکنیم. با فرض این که داده ها تصادفی هستند، تعداد رخ دادن رویداد مورد نظر را محاسبه می کنیم. اگر عدد به دست آمده به شکل چشمگیری از تعداد واقعی نمومه ها بیشتر باشد، آنگاه باید انتظار داشت هر چیزی که پیدا شده فریبنده باشد. البته این یک مسئله آماری ساختگی است تا یک مدرک برای پیدا کردن چیزی که به دنبال آن هستیم.

* 1. داده کاوی در اینترنت اشیاء

داده کاوی در اینترنت اشیاء جهت مدیریت مقدار زیاد داده هایی که توسط دستگاه های اینترنت اشیاء تولید می شوند، کاربرد دارد. داده کاوی شامل کشف و تجزیه و تحلیل دانش از مجموعه گسترده ای از داده ها است. هدف اصلی از استخراج داده، یافتن الگوهای مفید از مجموعه داده های بزرگ دریافت شده از دستگاه های اینترنت اشیاء (IoT) و سنسورها است. کشف دانش، تجزیه و تحلیل الگو ها و برداشت اطلاعات، مفاهیمی هستند که در بحث داده کاوی در اینترنت اشیاء استفاده می شوند. هدف اصلی از داده کاوی، ایجاد یک مدل کارآمد و توصیفی است که برای مجموعه داده مورد نظر مناسب باشد.

* + 1. k نزدیک ترین همسایه [[2]](#footnote-2)

الگوريتم k نزدیک ترین همسایه بر اساس اين اصل که داده هاي نزديک به هم در يک ديتاست داراي شباهت در خصوصيات مي باشند کار مي کند. اگر نمونه ها داراي برجسب باشند داده هاي بدون برچسب را مي توان در کلاس نزديک ترين داده به آن دسته بندي کرد. k نزدیک ترین همسایه با بررسي k عنصر داده نزديک به داده مورد سوال آن داده را در بيشترين تکرار کلاس در k نمونه دسته بندي مي کند.

به صورت عمومي، هر نمونه را مي توان يک نقطه در يک فضاي n بعدي در نظر گرفت که هر کدام از ابعاد توصيف کننده مقدار يکي از n ويژگي هاي نمونه مي باشد. مکان هر نقطه در فضا به اندازه فاصله نسبي آن از ديگر نقاط داراي اهميت نمي باشد. هدف کمينه سازي فاصله ميان نمونه هاي يک کلاس و بيشينه سازي فاصله ما بين نمونه هاي موجود در کلاس هاي مختلف است.

قدرت روش k نزدیک ترین همسایه در کاربردهاي واقعي بسياري اثبات شده است اما استثناهايي در رابطه با کارکرد مناسب k نزدیک ترین همسایه وجود دارد :

- روش k نزدیک ترین همسایه به فضاي ذخيره سازي زيادي نياز دارد.

- اين روش به تابع تشخيص شباهتي که استفاده شده حساس است.

- مشکل در تعيين مقدار مناسب براي k

- هزينه محاسباتي بالا

اوکاموتو و یوگامی دقت مورد انتظار k نزدیک ترین همسایه را به عنوان تابعي از ويژگي هاي دامين مورد استفاده مانند تعداد نمونه هاي موجود در مجموعه آموزش و تعداد خصوصيات مرتبط و نامرتبط، نرخ نويز، احتمال مربوط به هر خصوصيت و مقدار k نشان داده اند [3].

همانگونه که در قبل اشاره شد يکي از معايب روش هاي مبتني بر نمونه ميزان هزينه محاسباتي بالا در آن ها مي باشد. يک مسئله کليدي در بسياري از کاربردها تشخيص ويژگي هايي که در مدل سازي بايد استفاده شود توسط روش هاي انتخاب ويژگي تعيين مي شود [4]. انتخاب ويژگي مي تواند به بهبود دقت و مقياس پذيري و همچنين کاهش زمان دسته بندي کمک کند. مساله ديگر انتخاب نمونه هايي که براي مدل سازي بايد استفاده شوند مي باشد [5].

* 1. نتیجه گیری

فصل سوم: مروری بر تحقیقات اخیر انجام شده

3-1- مقدمه

در فصل قبل به چالش­های که در داده کاوی وجود دارد پرداخته شد. در این فصل به تحقیق­های جدیدی که در این زمینه انجام داده شده است، پرداخته خواهد شد.

3-2- مروری بر ادبیات موضوع

3-3- نتیجه گیری

در این فصل به تحقیقات جدید انجام شده در زمینه داده کاوی پزشکی پرداخته شد. با بررسی کارهای انجام شده در این زمینه می توان دریافت که در سال های اخیر به روش های ترکیبی داده کاوی از جمله شبکه عصبی فازی، دسته بندی مشارکتی، دسته بندی تلفیقی و... توجه بیشتری شده است. برخی روش ها مانند ماشین بردار پشتیبان کاربرد بیشتری در داده کاوی پزشکی داشته و نتایج بهتری را نسبت به سایر روش ها ارائه می دهد

مراجع

مراجع

[1] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction To Data Mining*. Boston: Pearson Addison Wesley, 2005.

[2] S. Sumathi and S.N. Sivanandam: *Data Warehousing, Data Mining, and Olap*, Studies in Computational Intelligence (SCI) **29**, 21–73 (2006).

[3] N. Y. Seishi Okamoto, “Effects of domain characteristics on instance-based learning algrithims,” *Theor. Comput Sci*, vol. 298, pp. 207–233, 2003.

[4] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[5] J. S. Sánchez, R. Barandela, and F. J. Ferri, “On filtering the training prototypes in nearest neighbour classification,” in *Topics in Artificial Intelligence*, Springer, 2002, pp. 239–248.

[6] Shobanadevi, A., & Maragatham, G. *Data mining techniques for IoT and big data - A survey*. In Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017 pp. 607–610. (2018).

1. Data Mining [↑](#footnote-ref-1)
2. K-Nearest Neighbor [↑](#footnote-ref-2)