



ارائه ی یک الگوریتم خوشه بندی ترکیبی با استفاده از k- میانگین هارمونیک و مراکز فازی

سینا دامی^۱، حسین داوطلب محمودی^۲

^۱استادیار، دانشگاه آزاد اسلامی واحد تهران غرب، گروه کامپیوتر، تهران، ایران dami@wtiau.ac.ir
^۲دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران غرب، گروه کامپیوتر، تهران، ایران hd.mahmoodi@yahoo.com

چکیده

در داده کاوی شیوه ای که اغلب برای پردازش داده استفاده می شود، خوشه بندی است. یکی از مشکلات مطرح در الگوریتم های خوشه بندی، عدم سازگاری نتایج با توجه به مقداردهی اولیه برای تعداد خوشه ها است. الگوریتم خوشه بندی k- میانگین هارمونیک به منظور رفع این مشکل بر روی داده های عددی به خوبی عمل می کند. با این وجود همچنان بر روی داده های ترکیبی با ویژگی های رسته ای عملکرد خوبی از خود نشان نمی دهد. در این مقاله روش خوشه بندی k- هارمونیک که واریانسی از شیوه k- میانگین معروف است، بر روی داده های ترکیبی همراه با تعاریفی از مراکز فازی خوشه و مقیاس فاصله ای ارائه شده است که هر دوی آن ها با تابع هزینه الگوریتم خوشه بندی k- میانگین هارمونیک در الگوریتم پیشنهادی استفاده می شوند. نتیجه آزمایش ها روی مجموعه داده های رسته ای خالص و ترکیبی نشان می دهد که الگوریتم پیشنهادی به خوبی روی مقداردهی تصادفی و یا غیر دقیق مرکز خوشه، عملکرد بهتری در مقایسه با k- میانگین هارمونیک دارد. حتی در حالتی که خوشه های تولید شده ویژگی های ترکیبی و متفاوت داشته باشند، این روش موثرتر بوده و نتایج خوشه بندی بهتری را تولید می کند.

واژه های کلیدی

الگوریتم خوشه بندی، k- میانگین هارمونیک، خوشه بندی فازی، مراکز فازی، داده های ترکیبی

۱- مقدمه

خوشه بندی داده یک تکنیک داده کاوی است که این امکان را فراهم می کند تا واحدهای اطلاعاتی با مشخصه های یکسان گروه بندی شوند بدین ترتیب اطلاعات موجود برای پردازش های آتی آماده خواهند بود [1]. به عبارت دیگر می توان گفت که الگوریتم های سریع و با کیفیت خوشه بندی نقش مهمی را در فراهم کردن مکانیسم های جستجو و پیمایش، به همراه سازماندهی حجم وسیع اطلاعات درون واحدهای معنادر خوشه ها، ایفا می کنند [2]. در گذشته خوشه بندی اسناد متنی، تنها برای بهبود دقت یا بهبود فراخوانی، در سیستم های بازیابی اطلاعات استفاده می شد [3]. اما در سال های اخیر کارهای انجام شده بر روی خوشه بندی به منظور مروری روی مجموعه اسناد [4] یا روی سازماندهی نتیجه های بدست آمده از موتور جستجو در پاسخ به پرس و جوی کاربر [5] پیشنهاد شده است. روش های

متفاوتی برای مقداردهی مرکز اولیه خوشه از قبیل خوشه بندی افرازی^۱ و سلسله مراتبی^۲ به منظور گروه بندی اطلاعات بکار گرفته شده اند [6]. روش k- میانگین^۳ (از نوع افرازی) ارائه شده توسط مک کوئین [7]، یک الگوریتم کلاسیک برای تحلیل خوشه است که با افراز واحدهای اطلاعاتی درون خوشه ها طبق معیار شباهت گام مهمی در گروه بندی اطلاعات ایفا کرده است. اما به دلیل حساس بودن به مرکز خوشه اولیه در موارد مختلف، نتایج ثابتی را ارائه نمی دهد [8]. در این مقاله سعی شده به کمک الگوریتم k- هارمونیک^۴ - (یک پروسه تعاملی مبتنی بر مرکز خوشه) به شیوه ای خاص با استفاده از مراکز فازی خوشه و با اعمال یک تابع هزینه خاص بر این مشکل چیره شود. تفاوت اصلی آن با روش k- میانگین، در این است که از جمع فاصله متوسط هارمونیک بین نقاط داده و مرکز خوشه استفاده می کند.

۲- کارهای مرتبط

در این بخش الگوریتم های خوشه بندی سلسله مراتبی و افرازی را برای مجموعه داده های ترکیبی مورد بحث قرار می دهیم. کاپیس و همکاران [۸] خوشه بندی سلسله مراتبی تجمیعی بالارونده^۵ را پیشنهاد دادند که هر داده اطلاعاتی را به عنوان یک خوشه منفرد در پایین ترین سطح در نظر می گیرد. ادغام خوشه ها طبق معیار شباهت تا رسیدن به شرط توقف ادامه می یابد. معمولاً یک ماتریس فاصله بین تمام خوشه ها تعریف می شود و هر دو خوشه نزدیک به هم ادغام می شوند تا کار خاتمه یابد بدین ترتیب یک نمودار دو بعدی از خوشه ها (دندوگرام^۶) ساخته می شود. با توجه به اینکه در این شیوه نیازی به دانستن تعداد خوشه ها نیست، عدم به روزرسانی خوشه ها از معایب آن به شمار می رود. در مقابل این تکنیک - های خوشه بندی سلسله مراتبی، تکنیک های خوشه بندی افرازی روی نقاط داده وجود دارند. اگر k تعداد خوشه های مورد نظر باشد، این تکنیک تمام k خوشه را در لحظه ایجاد

¹ Partitional

² Hierarchical

³ K-Means

⁴ K-Harmonic

⁵ Cumulative climbing

⁶ Dendrogram