بهینه‌سازی الگوریتم t-SNE با استفاده از الگوریتم بهینه‌ساز شیر مورچه

زهرا آهون[۱]، محمدرضا محمدرضایی[۲] و مهناز رفیعی[۳]
[۱]گروه کامپیوتر، موسسه آموزش عالی غیر انتفاعی اروندان خرمشهر، z_ahoon@yahoo.com
[۲]گروه کامپیوتر، واحد رامهرمز، دانشگاه آزاد اسلامی، رامهرمز، ایران، mohammadrezaei@iauramhormoz.ac.ir
[۳]گروه کامپیوتر، واحد رامهرمز، دانشگاه آزاد اسلامی، رامهرمز، ایران، m.rafir@srbiau.ac.ir

چکیده – در این مقاله یک روش جدید برای بهینه‌سازی t-SNE با استفاده از الگوریتم شیر مورچه ارائه‌شده است. نتایج در کارهای پیشین نشان می‌دهد که استفاده از این الگوریتم با هسته گرادیان منجر به از دست رفتن بخشی از اطلاعات می‌شود و ما به دنبال این هستیم که با استفاده از یک تابع بهینه‌سازی میزان از دست رفتن اطلاعات را کاهش و نتایج را بهبود دهیم. در این پژوهش هسته t-SNE که با استفاده از تابع گرادیان کار می‌کند با الگوریتم بهینه‌ساز شیر مورچه جایگزین شده است. الگوریتم شیر مورچه که یک الگوریتم بهینه‌سازی قوی می‌باشد باعث شده تا هنگام کاهش ابعاد داده‌ها بار اطلاعاتی حفظ‌شده و اطلاعات کمتری از بین برود. در کارهای آینده جهت مشاهده عملکرد روش‌پیشنهادی، نتایج به‌دست‌آمده از پیاده‌سازی‌ها با سایر الگوریتم‌های انجام شده پیشین در این زمینه مقایسه خواهد شد که نشان از بهبود عملکرد الگوریتم t-SNE با هسته جدید را دارد.

کلید واژه- ابعاد داده‌ها، برازندگی، شیرمورچه، کلان داده‌ها.

۱- مقدمه

یکی از مباحث مهم و مطرح در دنیای امروز مسئله ذخیره‌سازی، پردازش و تفسیر کلان داده‌ها[۱] است. کلان داده، به داده‌هایی گفته می‌شود که مدیریت و پردازش آن‌ها خارج از توانایی راه‌حل‌ها و سیستم‌های سنتی موجود است. حدود ۹۰ درصد کل داده‌های جهان، در چند سال گذشته تولید شده‌اند (عکس، ویدئو، حرکت ماوس، لایک‌ها و موارد مشابه دیگر، نمونه‌هایی از این داده‌ها هستند). همچنین تنوع داده و افزایش داده‌های غیر ساخت‌یافته باعث می‌شود که شرکت‌های بزرگی همچون گوگل و یاهو با حجم و تنوع بسیار زیادی از داده‌هایی که کاربرانشان تولید می‌کنند روبه‌رو شوند. ذخیره این حجم بالا از داده با تنوع زیاد بر روی کامپیوترها و ماشین‌های ارزان‌قیمت و ابزارهایی مانند اوراکل و موارد مشابه امکان‌پذیر نیست. یکی از کارهای اولیه که در این زمینه پیشنهاد می‌شود، فشرده‌سازی داده‌ها است. این امر در داده‌های بزرگ چندان کارساز نیست، زیرا یکی دیگر از خصوصیات داده‌های بزرگ، تنوع آن‌ها است[۱]. این داده‌ها از انواع مختلف داده‌ها تشکیل شده‌اند که این غیر ساخت‌یافتگی، فشرده‌سازی آن‌ها را دشوار می‌کند، به همین دلیل فشرده‌سازی، کاربردی در پردازش داده‌های بزرگ ندارد. ازآنجایی‌که پردازش و ذخیره‌سازی داده‌ها کار بسیار مشکلی است درنتیجه برای روبه‌رو شدن با این چالش جدید دانشمندان به دنبال راه‌حل و ابزارهای جدیدی جهت مدیریت و پردازش کلان داده‌ها هستند[۲]. یکی از این ابزارها و محبوب‌ترین آن‌ها هادوپ[۲] نام دارد. هادوپ را می‌توان حتی بر روی سیستم‌های ارزان‌قیمت و معمولی نصب و اجرا کرد و دیگر نیازی به کامپیوترهای گران‌قیمت و غول‌آسا نیست و می‌توان با شبکه کردن چند کامپیوتر معمولی و تقسیم داده‌ها بر روی این کامپیوترها، کلان داده‌ها را مدیریت کرد[۳]. در حالت کلی کلان داده یکی از واژه‌هایی است که برای معرفی Big Data مورداستفاده قرار می‌گیرد، البته واژه‌های معادل دیگری نیز ازجمله داده‌های انبوه، داده‌های حجیم، داده‌های عظیم، بزرگ داده، داده‌های بزرگ و موارد مشابه دیگر نیز استفاده می‌شوند. مفهوم کلان داده به معنی یک مجموعه داده در حال رشد است، به‌طوری‌که مدیریت آن با استفاده از مفاهیم و ابزارهای مدیریتی پایگاه داده‌های موجود دشوار است. این دشواری می‌تواند با ثبت،